

Foresight comes hurriedly of age

British plans to plot a strategy for research by untested iterative consultation are jeopardized by haste and are potentially flawed by the appearance of consensus.

THIS is a big week for British science — and its future. Hot from a flying visit to the United States, the science minister, Mr William Waldegrave, will be stumping the United Kingdom canvassing support for his plan to form research strategy by an untried technique called “foresight”. The principle is unexceptionable. Ask groups of well-informed people where they think efforts should be concentrated, put the opinions together and ask people for their opinions of the outcome, and then iterate the process until the outcome seems to represent consensus. Under the banner of “Delphi”, a technique much like this was used in the 1960s in attempts to define viable options in US strategic defence policy. But Britain intends that “foresight” should be an institutionalized and permanent instrument of public policy.

Having made wealth creation the central goal of this summer's white paper on research, and foresight its chief technique for making decisions, Waldegrave has no choice but to give it his blessing. In any case, the British Civil Service, adept as it is at turning ideas into committee structures, would not let him forget his promise. A steering group for what is called the Technology Foresight Programme is already in being, under the wing of the newly formed Council for Science and Technology (see *Nature* 365, 593; 1993). The next step is to recruit members of “expert panels” and then to compile the names of those from whom the panels will bounce questionnaires. To be fair, the documents circulated at Waldegrave's public meetings promise that the government is “anxious to listen” to doubts and reservations about the scheme. Here are some.

First, the process (whatever its promise) is being rushed unwisely. The steering group is under pressure to have all the machinery in place in the next few weeks. The process of defining national goals for British research will then be completed by the end of 1995. Thereafter, it seems inevitable that research-grant applications to the new research councils still being formed will be judged against that check-list. (Presumably particle physics and astronomy, exempted from the rigours of wealth creation in the white paper, will carry on as at present until somebody misguidedly raises the utilitarian question of why Britain should support any research expected not to contribute to national wealth.) It would be far wiser to spend 1994 on a careful trial of the new technique in part of the whole vast canvas of research — say the whole field of molecular biology and its applications (where hardly anything can go wrong).

Elsewhere, there is ample scope for disaster. One objec-

tive of the foresight exercise is to unite industry and the research community in the common definition of research objectives. But the industries concerned are the very same that have for decades been slow to see the commercial potential of technical innovation. Will their delphic opinions of research strategy be that funds should be spent on making up for the opportunities they have squandered? And if the consensus comes out against them, will their representatives then complain that vital national interests are being neglected? And who will then decide what will have become essentially political issues?

Part of the trouble is that there is no definition of what constitutes a meaningful research goal. The documents being distributed for this week's meeting speak of the identification of “generic technologies”, but say nothing of their conceptual significance. But what is a meaningful research goal? Would “quantum optics” qualify, on the grounds that a general improvement of competence in this field is likely to win benefits in fields as different as chip design and communications? Or will the allowed objectives be more specific, such as “light-emitting diodes in optical communications” (in which case the research community will become an extension of industry's development laboratories)?

A further difficulty is that the partition of research funds (the amount of which will be defined only in the budget due next month) between different fields is likely to be prejudiced by the internal structure of the foresight exercise. The expert panels will have their own fields of interest. However clear the consensus emerging from their activities, the proper division of resources between them will remain undecided? How will that be settled? By another delphic exercise? In the light of the numbers of people already working in these fields? Or by holding a moistened finger up to see which way the wind is blowing? If the British research community is not to be further demoralized, it would be proper to give weight to the second of these methods. An element of the third seems unavoidable.

With all that said, the research community should not lightly dismiss Waldegrave's many invitations this week to participate in the process. For the government has the research community over a barrel. Those who fail during 1994 to extol the potential for wealth creation of their own fields may find them omitted from the check-list guiding the distribution of funds in 1995. With only a little luck, the foresight process by itself will create a better understanding



between some British research and industry, but it is a coercive as well as a democratic process. If it goes badly wrong, it could finish off research in Britain — except perhaps in particle physics and astronomy. □

Modest growth for NIH

Congress's conference on the budget for the US National Institutes of Health has agreed on a 6 per cent rise.

FOR more than two decades, the White House (both Republican and Democratic) has consistently proposed budgets for the US National Institutes of Health (NIH) that it well knew were below what the Congress would eventually agree to spend. It was a silly game, but allowed members of Congress to look good in their home districts, where they could boast that they had voted to increase funding for vital (and politically popular) biomedical research. But this year it looked as if the game was up. The administration's budget request would have meant that most NIH institutes and divisions would have less to spend in real dollars.

Now, the conference of the two NIH appropriations committees has agreed to let NIH have an overall increase of some \$600 million, for a total budget close to \$10.9 billion — roughly 6 per cent above the final figure for the year just ended. In addition, it seems, both NIH committees have agreed not to repeat last year's fiasco in which earmarking funds to meet the demands of special interest groups played havoc with the budget process; feminist lobbyists managed to get \$210 million for breast cancer research added to the budget of the Department of Defense and a small biotechnology company's lobbyist won a trial of an AIDS vaccine that could not compete successfully in the peer review process. Let us hope that agreement sticks.

But will the 6 per cent suffice? The increase will not be met by jubilation in the US biomedical community and its professional organizations, which are forever telling the Congress of the virtues of research. Compared with past increases of 15 per cent or more, it does indeed seem meagre. Indeed, in a \$10-plus billion enterprise, \$600 million does look a little like petty cash. But these may be exceptional times. Bent on reducing the federal deficit, Congress has frozen discretionary spending (of which all scientific research is a part) for the next five years. So an increase for one agency must be at the expense of competing and often equally worthy causes.

Even so, this year may be the last in which biomedical research enjoys any special consideration. What will happen then? NIH (and its new director, Harold Varmus) should turn immediate attention to winning greater value from its budget. A review of the balance between intramural and extramural research is overdue. Then the peer-review system, which now ranks more than 20,000 grant applications a year, needs overhauling. But all steps in those directions will hurt not only researchers but also their institutions. Will Varmus become a super-lobbyist, not only in the Congress but in the board rooms of the biomedical corporations? □

Thatcher's tale

This year's most celebrated political autobiography seems not to be recognizable as history of science.

MRS Margaret (now Lady) Thatcher is the best-known of Britain's prime ministers since Sir Winston Churchill, and was the first to be qualified as a scientist. Even though she was forced out of office three years ago, she remains a powerful influence in British politics. Her declaration to a radio audience on Monday that her successor, Mr John Major, now deserves support because of his supposed recent conversion to "traditional values" may even help him in the public opinion polls. The broadcast was itself a contribution to the campaign to sell copies of her book called *The Downing Street Years* (HarperCollins, London; £25.00).

Inevitably, the chief interest of this work is the author's account of her often tempestuous relations with her colleagues, many of them still in politics. Touchingly, her personal stationery carrying a dedication (to husband, family and ex-helpers) now reproduced shows that, in her mind, her fellowship of the Royal Society still ranks alongside her Order of Merit and membership of the Privy Council, but researchers will search in vain for an explanation of why they and their institutions were dealt with so shabbily during her regime. Indeed, the author's only substantial account of an intervention in scientific affairs is of the circumstances leading to her endorsement of the view that global warming is potentially a threat to the global environment.

Surprisingly, Thatcher explains that her conversion to this cause derives from her recognition, some time before 1987, that too much public support was directed towards defence science, and too much towards "the development of products for the market rather than to pure science". She adds that "as someone with a scientific background, I knew that the greatest economic benefits had always resulted from advances in fundamental knowledge rather than the search for specific applications". She puts the case for basic research vividly: "transistors were not discovered by the entertainment industry seeking new ways of marketing pop music, but rather by people working on wave mechanics and solid-state physics". But these opinions do not accord with the recollections of those in basic research around 1987, nor, interestingly, with her successors' devotion to wealth creation.

For science as in many other ways, Thatcher is a greater disappointment with hindsight than she appeared at the time. Intelligent, energetic and determined, she might have done for British science what President François Mitterrand did for France about the same time. And if basic research is as valuable as she now says she knew it to be, why did she tolerate so much beastliness in her government's handling of research? Was not that the time when British academic institutions were under the most severe pressure or, as she now puts it, "by exerting financial pressure we had increased administrative efficiency and provoked overdue rationalisation"? Even in this peripheral neck of Thatcher's woods, the history is hardly recognizable in the text. □

Space station design falls hostage to US–Russia missile deal...

Washington. The design of the space station being planned by the National Aeronautics and Space Administration (NASA) seems likely to be shaped more by what the United States feels it must offer the Russians in return for their signature on the Missile Technology Control Regime — an agreement to stop the spread of weapons technology — than by engineering or scientific considerations.

A final design for the proposed 'international space station' being negotiated between the United States and Russia will not now be ready until 22 December, NASA administrator Dan Goldin told a congressional hearing last week. And according to congressional aides, it will not be finalized until Vice President Al Gore visits Moscow after the Russian elections in November.

Goldin was speaking after his return from a visit to the Russian capital that coincided with the bloody showdown on Moscow's streets. He said that he had been strongly impressed by the commitment of the Russian team which had negotiated with him on the details of the recent US–Russia space cooperation agreement (see *Nature* 365, 101; 1993) against a backdrop of gunfire.

But he told a House of Representatives space subcommittee hearing that he had not yet discussed the space station, the third phase of the three-part deal, with the Russians. "Right now we've only talked about phases one and two", he said.

The space station is due to be discussed at another set of meetings this week, Goldin said. Under Gore's plan, the latest design for the space station will be presented to the White House by NASA on 1 November.

The first two phases of the US–Russian space deal are bilateral agreements. But the third phase is meant to involve the United States' existing space station partners — namely Japan, Europe and Canada — as well. NASA officials were in Paris last week briefing the European Space Agency on their progress with the Russians.

At an earlier hearing of the space subcommittee, John Gibbons, President Clinton's science adviser, was pressed on why he had no details of the plan. In reply, Gibbons said he did not expect to see an agreement as such on 1 November. "I expect to see a set of options", he said.

Goldin appeared to admit his frustration at the way the decision on the final space station design is being drawn out. "I do not have full control of this process", he said. "I will just deliver the data."

The lack of information on the agreement with the Russians has angered congressmen who fought to save the space

station from budget cuts this summer — eventually prevailing in the House by 216 votes to 215 — only to find it once again bogged down in yet another redesign.

Goldin says NASA is now working on two programmes. One is the space station Alpha as accepted by Clinton in the summer, the other is the 'international' station involving the Russians. Boeing has already been named as prime contractor; but it will not now be told what it has to build until next March.

In testimony from the Congressional Research Service, space specialist Marcia Smith described the Gore plan as "fundamentally a foreign policy decision, not a choice based on space policy". Smith said that merging the space station with foreign policy objectives was likely to "complicate, rather than complete" the programme.

Meanwhile, the technical problems resulting from the US–Russian deal continue to mount. For example, in order to reach a station repositioned to suit the Russians, shuttle launches from Florida would have to fit in a five minute 'window' each day.

Goldin and Gibbons do not consider this a problem. But James Sensenbrenner (Republican, Wisconsin) says that only 30 out

of 86 shuttle launch attempts so far would have coped with such a restriction.

More importantly, the death of the Advanced Solid Rocket Motor in Congress two weeks ago leaves the United States with a lower shuttle launch capacity than its earlier plans had assumed.

The biggest question mark hangs over the wisdom of allowing the fate of an international programme that has already been almost destroyed by the vagaries of the US Congress to depend on the even greater vagaries of Russian politics.

In reply to a question from Tim Roemer (Democrat, Indiana) on what would happen if the Russians breach the missile treaty at some future date, Goldin said: "We'll have a back up as part of our planning process." Congress is unconvinced: in the words of one aide, such a "twin track" policy is "absolutely not possible".

If this is the case, then the Clinton administration may have to wait for political stability in Russia before it can proceed with the space station. But station supporters in Congress are asking whether such stability is likely to be assured by the time they next have to justify funding for the project in nine months time.

Colin Macilwain

... as Europe plans solo comet trip

London. The European Space Agency (ESA) is expected to give its formal approval next month to plans for the launch in 2003 of the first spacecraft to land scientific instruments on the surface of a comet. US scientists are being invited to participate; but the project will not, as originally intended, be run jointly with the US National Aeronautics and Space Administration (NASA).

The so-called Rosetta mission — named after the complex hieroglyphs found on the Rosetta stone in Egypt — will send a spacecraft to fly alongside a comet, most likely Schwassam-Wachmann 3, as it passes through the Solar System and eventually swings around the Sun. It will form the third of four 'cornerstone' missions that have been planned as the basis of ESA's space science programme since these were agreed in the mid-1980s.

Original plans for the mission were drawn up jointly between scientists at ESA and NASA, with the costs of up to \$2 billion to be split on a 40:60 basis. Funding from the two agencies was to have supported a more ambitious mission, including a vehicle that would have returned samples of the comet's surface to Earth for analysis.

Last year, however, following indica-

tions from NASA that the US contribution was unlikely to materialize, ESA agreed that the sample-return element was not essential to the mission's success, and that many of its scientific goals could be achieved by a more modest mission.

It was also accepted by ESA that European development work on solar arrays meant that it would no longer be necessary to rely on US technology to power the space

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

**Rosetta: a 'cornerstone' mission
(artist's impression).**

European Space Agency

vehicle during its initial encounter with the comet while at a long distance from the Sun.

The mission has been redefined without the sample-return element, reducing its costs to an estimated 600 million units of account (US\$720 million). But its technical feasibility, scientific value and reduced cost are now such that European scientists are confident the mission can be accomplished without the aid of NASA.

As a result, ESA's space science advisory committee decided at a meeting last month to recommend an earlier than anticipated launch date of 2003, with the initial comet rendezvous scheduled for 2008.

The main space vehicle will travel in parallel with the comet until the two reach the comet's perihelion in 2011. Proposals are now being discussed with scientific groups and funding agencies in various countries — including the United States, Russia and Japan, as well as ESA member states — to build instruments for a separate vehicle that would be dropped on to the comet's surface to make physical and chemical measurements.

Planetary scientists are pleased with the priority now being given to the Rosetta mission. "I am very excited by the prospects", says Eberhardt Grün of the Max Planck Institute for Nuclear Physics in Heidelberg, who has been working on plans for the mission for almost ten years.

US scientists, frustrated by NASA's decision last year to drop funding for its own proposed comet rendezvous mission, CRAF, are also looking forward to the opportunity they are being offered to participate in a cometary mission, even one in which Europe will now be setting the timetable.

"For example, an instrument developed for the CRAF mission designed to send back images of the surface could now be used on Rosetta", says Thomas Ahrens of the Seismological Laboratory at the California Institute of Technology, one of the main authors, with Grün, of a 1987 report outlining the scientific goals of the mission.

The main disappointment will come from European astronomers, however, who had been hoping that the third cornerstone launch slot would be given to the Far Infrared Submillimetre Wave Telescope, FIRST (the first two slots are for the SOHO/Cluster space vehicles, and the XMM telescope).

After a presentation of both Rosetta and FIRST, ESA's advisory committee decided at last month's meeting to recommend that the Rosetta mission be launched first, in which case the telescope will not be launched until at least 2006. One factor said to have influenced the committee's decision is that there are a number of astronomy missions already waiting to be launched.

Both decisions are expected to be endorsed by ESA's science programme committee when it meets in Paris on 4 November to discuss the advisory committee's recommendations.

David Dickson

Italy cuts research funds but adds university jobs

Rome. The Italian government is to create thousands of new university teaching and research posts to meet rising student demand, even though it is planning to cut the national research budget by 3 per cent — and is recommending even larger cuts in its direct funding of university research.

The cuts are being imposed as part of a package of stringent measures intended to deal with the country's economic crisis. Nevertheless, Umberto Colombo, the research minister, has managed to implement a decision made as part of a previously agreed three year plan for 1991–93 (but blocked by bureaucracy) to increase the number of permanent university staff.

As a result, funding for 1,250 full and associate professors, and for 2,000 research staff, will be allocated to universities by the end of the year. Discussions are taking place between the Ministry of Universities and Research, the National Research Council and universities to decide the precise number of new posts to be given to individual institutions.

In a separate move, universities and research institutes were told two weeks ago that they are to be exempted from the government's freeze on all recruitment to civil servant posts, allowing them to fill existing vacancies.

Both measures have been prompted by the fact that the number of Italian university students has tripled in the last 25 years, to a total of 1.5 million. But over the same period, there has been no accompanying increase in teaching staff; the number of professors, for example, has remained essentially unaltered at 45,000.

Despite staffing improvements, the government is recommending that the funds universities devote next year to research should be reduced by 15 per cent on 1993. Colombo has long criticized universities for using these funds in an unfocused way without adequate targeting (see *Nature* 364, 473; 1993).

But Colombo also points out that he is now giving universities more control over

Research institutes facing funding cuts

	1993	1994
National Research Council	1110	1050
Italian Space Agency	850	800
Nuclear and Alternative Energy Agency	600	550
National Institute for Nuclear Physics	440	400

All figures in billions of lire.

the way they allocate their budgets, which will be provided by the government as a lump sum. This means that they can choose to increase the government's suggested budget for basic research with funds saved from other administrative areas.

The overall research budget is likely to be at least 3 per cent lower than last year. The first round of cabinet debate on the budget has left the research ministry with much less than it had hoped for. As a result, says Colombo, "small sacrifices will have to be made in nearly all areas". But he says that, in comparison with other ministries, which have had to accept cuts averaging at about 6 per cent, research has not done too badly. The budget should receive final approval from the parliament by the end of the year.

Colombo also points out that he has managed to secure an increase in funds for capital investments in research facilities. These will allow for the completion on the national Synchrotron Radiation Facility in Trieste, and building of the national telescope Galileo.

Luciano Maiani, head of the National Institute for Nuclear Physics (INFN), says that the cuts hurt, but acknowledges that they have been made necessary by the economic situation. But he says that INFN will be able to absorb the cuts without a major impact on its research activities.

The INFN, which runs four institutes and 12 university research units with at least 800 scientists, is now in the first year of a five-year plan introduced by Maiani. His budget for next year has now been cut back by 10 per cent.

To meet the new budget targets, Maiani has delayed two long-term projects to construct powerful heavy-ion accelerators. But time has also been bought by continuing delays to the proposed Large Hadron Collider (LHC) at the European particle physics laboratory (CERN) in Geneva, now expected to be ready by 2001 if it is approved at CERN council this December.

As a result, INFN has been able to put back the construction of LHC detectors for which it is responsible. And Maiani says that he has been reassured by — even though he is not counting on — promises from the ministry that the 10 per cent cut will be made up over the following two years.

Alison Abbott



Research minister Colombo (top) and INFN director Maiani.

Weather contributes to record ozone loss

London. A combination of satellite, balloon and ground-based data have confirmed that this year's Antarctic ozone 'hole' is the deepest ever recorded. But weather conditions may have been as much to blame as the build-up of atmospheric chlorofluorocarbons (CFCs).

On 6 October, a balloon launched from the South Pole by the Climate Monitoring and Diagnostic Laboratory (CMDL) of the US National Oceanic and Atmospheric Administration found total ozone concentration of only 90 Dobson units (DU). This compares to pre-depletion values of 275 DU, and to last year's minimum of 105 DU.

On the same day, a surface-based Dobson spectrophotometer recorded an ozone concentration of only 88 DU. These measurements represent the lowest value of total column ozone ever measured on Earth. (A Dobson unit is a measure of the thickness of the ozone layer if it was brought to the Earth's surface.)

Preliminary data from the National Aeronautics and Space Administration (NASA)'s total ozone mapping spectrometer (TOMS) instrument aboard the Russian Meteor-3 satellite also show October ozone values of less than 100 DU in a region near the South Pole. And at Halley Station, scientists from the British Antarctic Survey observed a minimum daily mean ozone value of 100 DU during October, compared to 107 DU last year.

The depth of this year's hole has come as a surprise to most atmospheric scientists. Chlorine released in the stratosphere from man-made chemicals such as CFCs is believed to be the main cause of stratospheric ozone depletion, and there is certainly more in the atmosphere this year than last.

But this should not necessarily have led to a deeper hole, as chlorine is released into its active form primarily in the presence of heterogeneous surfaces on which key chemical reactions can take place. These surfaces can be provided by polar stratospheric clouds (PSCs), which form only under extremely cold conditions, or by aerosol particles which occur over a restricted height range.

Last year's destruction in Antarctica seemed to account for virtually all of the ozone present at the limited altitudes over which such surfaces can occur. As a result, most scientists expected the ozone depletion to have reached its maximum this year, in spite of the extra chlorine present.

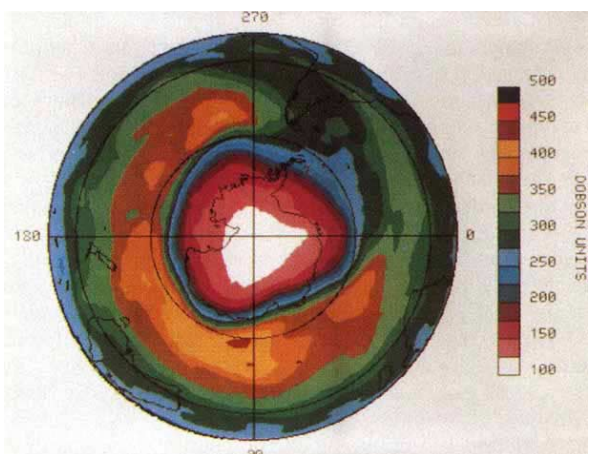
David Hofmann, from the CMDL, believes that the vertical resolution of the balloon-borne instruments may answer the question of why this year's ozone concentration is so low. During soundings over the South Pole on 6 and 12 October, the balloons recorded complete ozone destruction from about 13.5 to 19 km. But the data also show that most of this year's additional

depletion took place higher in the stratosphere than last year, from 18 to 23 km.

According to Hofmann, the balloons also detected record low temperatures in the Antarctic stratosphere. This means that PSCs may have formed at higher altitudes than is usually possible. Rich McPeters of NASA's Goddard Space Flight Center in Maryland says that this year's Antarctic vortex seems to have been very stable, which could have allowed unusually cold conditions to develop.

As a result, the unprecedented depth and vertical extent of this year's hole may owe more to meteorological conditions than to extra anthropogenic emissions of CFCs and other chlorine-containing species.

But both effects are likely to have played a role. Indeed the cold temperatures themselves may not be entirely free from anthropogenic influence. Hofmann points out that ozone acts as a stratospheric heater; ozone depletion can therefore lead to lower



This year's ozone hole covers roughly the same area as last year, but is much deeper (image taken 6 October).

stratospheric temperatures, and hence more depletion, in a positive feedback loop.

It remains unclear how strong this feedback actually is. But Hofmann points out that, for the past five years, severe ozone destruction in the Antarctic has been accompanied by relatively stable vortices. Thus ozone destruction might even play a role in determining meteorological conditions such as those seen this year. **Gabrielle Walker**

National supercomputer centre urged for US

Washington. The United States should set up a national, multi-agency supercomputer facility in order to push back the limits of high-performance computing, according to a panel of experts set up by the National Science Foundation (NSF).

Lewis Branscomb, formerly Research Director at IBM, now at Harvard University's Kennedy School of Government, says that such a facility would sit at the apex of a "high-performance computing pyramid" serving all the nation's research needs. He estimates that it would cost about \$50 million a year to set up and run.

Several members of the NSF panel dissented from the proposal to set up a national centre, arguing that it would be both ineffective and expensive. But Branscomb, a physicist who headed the National Bureau of Standards under President Jimmy Carter, argues strongly for "one national facility that would stretch the [supercomputer] industry".

Establishing a supercomputer pyramid would allow four existing supercomputer centres to be retained, alleviating fears that some or all of them might be closed. And the panel also additionally recommends that NSF should help every research university in the United States to buy its own mid-sized supercomputer, using the new 'massively parallel' architectures, over the next five

years. Support for scientists and engineers buying their own workstations should be doubled, it says.

According to Branscomb, the panel is confident that the additional funding required to fulfil its recommendations will become available under the federal High Performance Computing programme. This was launched under former President George Bush, and remains strongly supported by the Clinton administration.

The report says that a new central facility would provide "significant stimulus to commercial development" of future supercomputers delivering teraflop performance, that is, that measured in a million million floating point operations per second. It suggests that the centre should be run jointly by NSF and other federal agencies.

Although urging that the four existing centres should be retained, the report also calls on them to exercise more flexibility and to work more closely with other institutions. "The centres should no longer be seen just as places where gigabytes are farmed out to scientists," Branscomb says.

The NSF, whose new director Neal Lane was sworn in last week by President Bill Clinton, will prepare a plan for implementing the panel's recommendations for submission to the foundation's governing board early next year. **Colin Macilwain**

Biotechnology rules: Spain falls into line

Barcelona. Spain's first biotechnology bill is to be debated by Parliament this autumn. It is expected to provoke controversy, as it fails to give either public interest groups or representatives from Spain's 17 autonomous regions any say in risk control.

The bill also fails to allocate responsibility for accidents to manufacturers of genetically modified products. As a result, some critics fear that farmers or other users of genetically manipulated organisms could be personally sued if things go wrong.

The bill is an attempt to incorporate two European Communities (EC) directives on the use of genetically manipulated organisms into Spanish law.

A National Commission on Biosecurity (CNB) will be set up, whose members will

include expert scientists and representatives from six ministries: public works and transport (which is responsible for environmental affairs and which originally drafted the bill), health, agriculture, education and science, industry and the interior. However it excludes trade unions and environmental groups, contrary to the directives' recommendations.

The bill also establishes a new licensing authority for genetically modified organisms, made up of officials from the ministries of transport, health, agriculture and industry. This authority would have full executive powers, and would not be bound by the CNB's reports on risk.

Spain's autonomous regions will maintain their responsibilities for regulating the

confined use of genetically altered products. But the authority's ruling on products released into the environment will apply nationally — another potential point of contention when it is debated in Parliament.

The exclusion of the autonomous regions from both the licensing authority and the CNB is certain to generate criticism, particularly from Catalonia. In 1990, problems encountered with an experiment involving genetically engineered plants carried out by the Catalan government's Institute of Agriculture — including the fact that the crop spread beyond its confines — meant that the whole crop had to be burnt.

But the exclusion of the autonomous regions may be intended to allow the socialist government, which does not have a full majority in the national Parliament, to negotiate a consultative role for local governments within a broader package of political agreements.

Daniel Borillo, a lawyer specializing in gene laws, claims that the exclusion of many interest groups is contrary to the spirit and letter of the EC directives. He also criticizes the bill for failing to allocate responsibility for accidents.

Borillo says that the bill should have imposed an obligatory insurance on manufacturers to cover the risks of their products, as in other EC countries. But he claims that the government has bowed to pressure from industry to protect its operating costs.

Many scientists are also unhappy that the bill was drafted behind closed ministry doors. Emilio Munõz, former president of CSIC, Spain's main research organization, and now based at Madrid's Institute of Social Studies, says that a wider public debate is needed.

Luis Angel Fernandez

German changes face opposition

Munich. Germany's Social Democrats are threatening to block the passage of a proposed law revising the regulations covering genetic engineering when the law, which was approved by the country's federal Parliament (Bundestag) two weeks ago, is debated in its lower house, the Bundesrat, next month.

In particular, the Social Democrats — who are in the minority in the Bundestag, which is dominated by the Christian Democrats, but have a majority in the Bundesrat, made up of representatives from the individual *Länder* — object to a newly added clause that would allow field experiments with genetically modified organisms without a public hearing.

The main aim of the new bill is to streamline current procedures for approving all genetic engineering experiments. It is being actively supported by the country's chemical and pharmaceutical industries (as well as by many research organizations) which argue that excessively rigid legislation is forcing companies to locate research and development activities in foreign countries.

According to the law put to the Bundestag, there would be a substantial reduction in the amount of paperwork required for experiments considered to present no risk or low risk. In addition, the obligatory two-month waiting period between registering a low-risk experiment and starting experiments would be halved (see *Nature* 359, 93; 1992).

Before the debate in the Bundestag, the Social Democrats had said that they were willing to accept the main changes contained in the bill, which had been under detailed technical discussion between scientists and politicians for nearly a year, but that they did not want to reduce the waiting period.

Only a few days before the bill appeared

in Parliament, however, the Christian Democrats added a further clause, which had not been discussed with the opposition parties, allowing the release into the environment of genetically modified plants without a public hearing.

This sudden addition, which has been described by environmentalist groups as a "scandalously cloak and dagger" move, caused most Social Democrats to abstain from the final vote in the Bundestag, and two to vote against the bill.

If the Social Democrats persuade the Bundesrat to block the bill next month, it will go to a mediation committee between the two legislative bodies to reach a compromise. And if this happens, the revised law is unlikely to be passed, as hoped, by the end of this year.

Alison Abbott

UK report criticizes EC directives

London. Britain's biotechnology industry has persuaded a committee of the House of Lords to back its claims that excessive regulation, much of it emanating from the European Commission in Brussels, is reducing its ability to compete effectively with companies in both the United States and Japan.

In a report published in London last week, the Lords Select Committee on Science and Technology delivered a strong attack on two directives issued by Brussels in 1990, one on the contained use of genetically manipulated organisms and the other on their controlled release, which are now required to form the basis of legislation in all member states.

The committee claims that both directives are unscientific in their content, impose an excessive burden on the research community and biotechnology companies, and fail to take into account critical com-

ments made at the time they were drafted.

It calls for the government to push for amendments to the directives, making them less onerous on industry, and argues strongly against the idea that socio-economic need — the so-called 'fourth hurdle' — be taken into account when assessing new safety regulations.

Unsurprisingly, the report has received a warm welcome in the biotechnology industry. In contrast, there has been a more lukewarm response from those responsible for implementing the existing regulations.

John Beringer, for example, professor of molecular genetics at the University of Bristol, and chairman of the Advisory Committee on Releases to the Environment, points out that a number of steps have recently been introduced to streamline the application procedures for experiments and reduce their burden on researchers. □

Nobel rewards two laboratory revolutions

London. This year's Nobel Prize for Chemistry has been awarded to Michael Smith for his work on site-directed mutagenesis, and Kary B. Mullis for the invention of the polymerase chain reaction (PCR). Both techniques have already revolutionized the procedures used in research laboratories and biopharmaceutical companies.

Site-directed mutagenesis uses mismatched synthetic DNA fragments (oligonucleotides) to alter specific nucleotides along a sequence of DNA. This produces predictable changes in the amino-acid sequence of the translated protein, and makes it possible both to work out the how the different parts of a protein contribute to its function, and eventually to design proteins. PCR also uses oligonucleotides, but in a cycle that lets an extremely small amount of DNA be amplified exponentially. Site-directed mutagenesis and PCR are often combined.

"We were in a unique position to develop site-directed mutagenesis", says Smith, who works at the University of British Columbia in Canada, referring to work carried out with his colleagues. "We had information on duplex stability with mismatches, and at the time one had to be a chemist to make oligonucleotides."

"Few genomic sequences were known, but I was on sabbatical in Fred Sanger's lab, and so had the sequence of the bacteriophage Φ X-174, which was relatively easy to work with because it was single stranded DNA." The prize is "a great honour" he says, be-

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Michael Smith: "In a unique position to develop site-directed mutagenesis".

cause "there are so many scientists doing good things".

Kary Mullis, who is now a freelance consultant based in La Jolla, California, was stirred but not shaken by news of his prize: "From what people told me, I figured it would happen eventually, and it might as well be sooner rather than later."

Mullis says that the prize recognizes the significance of a finding and not the work of individuals during their lifetime. "Whether I'm a maverick, weirdo, or whatever, if the finding is important, then it should get the Nobel prize."

Alan Fersht, a 'protein engineer' who is professor of organic chemistry at the University of Cambridge, says he is delighted that gene technology had been honoured by the chemistry prize. "This is modern organic chemistry; for both techniques you need chemistry to make the oligonucleotides and enzymes to finish it off." He says that "site-directed mutagenesis has totally transformed my science".

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Kary B. Mullis: "If the finding's important, it should get the prize".

Peter Little, a molecular geneticist at Imperial College, London, describes Smith and Mullis as a "surprising pairing of winners", but adds that "there was no doubt that they should win the Nobel prize". He welcomes the award for PCR because "it has changed the way molecular biology is done", and made the public more aware of the possibilities of DNA analysis. He says that site-directed mutagenesis is "an incredibly important tool" and that "we have not yet had time to see its full effects on medicine".

Alec Jeffreys, who invented DNA-fingerprinting in his laboratory at the University of Leicester, has used PCR for many applications — including its first use to identify a murder victim. He says he now finds it impossible to conceive of a molecular genetics laboratory without PCR. "Like all good scientific ideas, it's exquisitely simple and, in retrospect, obvious", he says. "I'm sure that thousands of people like myself are kicking themselves that they didn't think of it."

Kimberley Carr

Prize lifts prospects for gravity wave detectors

London. The Nobel Prize for Physics has been awarded to Joseph Taylor and Russell Hulse of Princeton University for their discovery of binary pulsars. This is the second time that the physics prize has gone to research on pulsars — rotating neutron stars that send a lighthouse-like beam of radiation sweeping through space.

Pulsars were discovered in 1967 by Anthony Hewish of the University of Cambridge and his student Jocelyn Bell. Their observation of oscillations from pulsars at radio frequencies confirmed that when a star explodes in a supernova, electrons can fuse with protons in a highly dense, imploding core to create a compact body composed only of neutrons. Some neutron stars were predicted to have a strong magnetic field, which would cause them to emit intense beams of radiation at each magnetic pole. If the star rotates, the beam appears to turn on and off when viewed from Earth.

While the implications of this discovery lay primarily with stellar and nuclear physics, Taylor and Hulse's observation of a

pulsar (PSR1913+16) circulating a companion in a binary system confirmed an important prediction of Einstein's theory of general relativity: that a rotating mass should emit gravitational energy as waves in much the same way as a rotating electric charge emits electromagnetic radiation. This had never been shown experimentally, because the amount of energy emitted is so small.

In 1919, Arthur Eddington provided confirmation of Einstein's theory that a gravitational field represents a curvature of spacetime. Taking the opportunity offered by an eclipse, he observed that light rays from a star are "bent" by the Sun's gravitational field, as the theory predicted.

Taylor and Hulse reasoned that they should be able to detect gravitational waves from their binary pulsar because of the close orbits of the pair — as the binary loses orbital energy, the pulsar's orbit spirals inwards and the period decreases. By 1978, Taylor's group had confirmed that the orbit of PSR1913+16 decreased by the predicted

amount of 75 microseconds per year (within errors of 20 per cent).

It was this quantitative agreement (now with much tighter error margins) that provided the compelling test of the theory, says Clifford Will of Washington University, in St Louis, Missouri, who is a specialist on tests of general relativity. He says that the discovery of a binary pulsar was somewhat fortuitous, in that it would have been undetectable had the signal been slightly weaker. But he praises Hulse's persistence in pursuing his observations instead of rejecting them as instrumental artefacts.

Hulse says he was more disappointed than excited when he first noticed that the pulsar's period appeared to vary between observations. He says that with further observations, he became convinced that this variation was not an instrumental artefact, and that he eventually realized it was due to the presence of a companion body.

Although Taylor stresses that he and Hulse had no intention of "setting out to work on relativity", he explains that the

discovery of evidence that gravitational waves from astrophysical systems have observable consequences provided the impetus for proposed gravity-wave detectors, such as the proposed Laser Interferometer Gravitational Wave Observatory (LIGO) in the United States, and the Franco-Italian detector Virgo. Both are huge interferometric systems designed to detect gravitational waves produced from astrophysical events as they wash over Earth.

Hewish, whose discovery earned him the Nobel prize in 1974, says that Taylor and

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

**Joseph Taylor and Russell Hulse:
persistence rewarded.**

Hulse's prize should brighten the prospects for these expensive projects, which have fought for survival for years. LIGO's director Rochus Vogt of the California Institute of Technology says that the award is "an endorsement by the Nobel committee of the importance of gravity waves", and "will definitely help [LIGO] in the long run".

Meanwhile, Hulse himself switched to research on plasma fusion at Princeton soon after completing his thesis. One reason, he says, was the lack of long-term opportunities in astrophysics. **Philip Ball**

SSC faces last ditch challenge in Congress

Washington. Three hotly disputed research programmes were each given full funding at a joint meeting last week designed to iron out differences between representatives of the US House of Representatives and Senate.

The conference agreement on the Energy and Water budget bill allowed \$640 million for the Superconducting Super Collider, \$110 million for the Integral Fast Reactor (IFR) programme and \$36 million for the B Factory to be built at Stanford, California. The outcome means a certain reprieve for the IFR, which the House had voted to shut down (see *Nature* 365, 99; 1993), and that work on the B factory can start immediately.

But the funding for the SSC was still set to face a last gasp challenge in the House earlier this week. SSC supporters were confident that the challenge would fail for procedural reasons, despite the unpopularity of the project in the House. The entire budget, comprising 13 separate bills, is set to be agreed by today (Thursday).

Colin Macilwain

French gene laboratory gets a new lease of life

Paris. Uncertainty over the future of the Généthon laboratory near Paris has ended, following a decision by the French Muscular Dystrophy Association (AFM) to launch two long-term research programmes, one to identify disease genes, the other to work out what they do, and how.

Généthon was set up in 1990 jointly by AFM and the Centre d'Etudes du Polymorphisme Humain (CEPH). It is based at AFM's headquarters at Evry near Paris, and has built a reputation as one of Europe's most advanced gene mapping centres.

Earlier this year, however, AFM announced that it would not support further long-term research programmes after Généthon completed its mapping programme this year. It said that it would instead seek to eliminate bottlenecks in gene therapy research by funding research groups and biotechnology companies (see *Nature* 361, 671; 1993).

Under new plans, to be announced next month, Généthon II will identify genes using the laboratory's physical and genetic maps of the human genome. Its broader aim is to create a centre of excellence to develop improved positional cloning techniques to be made available to research groups worldwide. The programme will be headed by Jean Weissenbach, who published a 'second generation' genetic map of the human genome last year (see *Nature* 359, 380; 1992).

Axel Kahn, director of the INSERM Laboratory on Genetics and Molecular Pathology at the Cochin Institute of Molecular Genetics in Paris, will lead Généthon III. This will work out the role and function of genes once they have been identified. In particular, it will test what happens when genes of unknown function are inserted into mice, develop less-empirical ways of assessing gene function, and create animal models for particular diseases.

Généthon II will begin immediately. The start of Généthon III will depend on whether AFM reaches its target of FF400 million

(US\$71 million) in a television appeal in December (last year it raised FF314 million in this way). AFM plans eventually to back research on vectors for gene therapy (Généthon IV) and clinical trials (Généthon V).

Bernard Barateau, the president of AFM, says that several companies, including Rhône-Poulenc Rorer, are planning to invest in Généthon II and III. The local authority has also made available to AFM land next to Généthon to help it attract biotechnology companies and other research groups to the site.

Private organizations and charities contribute significantly to the French genome effort. The government is keen to integrate such organizations into its national strategy for genome research. François Fillon, the minister of higher education and research, is said to be impressed by Généthon's approach; Kahn says he has also expressed an interest in funding AFM's new programmes directly.

But although GREG funds some research at Généthon, close collaboration between the government-backed programme and AFM has been blocked by a long-standing deadlock between Barateau and Piotr Slominski, the director of GREG. One solution may be that proposed by Fillon to appoint one research organization to take responsibility for co-ordinating each area of biological research.

It is not known if Fillon intends to include genome research in this scheme. But François Kourilsky, director general of the Centre National de la Recherche Scientifique (CNRS), says that CNRS could become the lead agency for genome research. If this were accepted, he says, he would bring together the opposing parties by "imposing fortnightly meetings".

Meanwhile, Slominski is taking comfort from the fact that Fillon has increased GREG's budget for next year. "This clearly shows his willingness to develop genome research", he says. **Declan Butler**

Bid to launch 'European PhD'

London. Scientists from the 12 member states of the European Communities (EC), as well as four non-EC European states, have set up an executive board to oversee the development of a European doctorate in biotechnology. A meeting in Luxembourg last week discussed how the new doctorate could be awarded to those completing a PhD programme recognized throughout Europe, and was told that it might form a prototype for similar programmes in other disciplines.

The initiative has been backed by Antonio Ruberti, the European Commissioner for research, who told the Luxembourg meeting that it was in line with the commission's attempts to increase the mobility of doctoral and postdoctoral fellows within the EC. In addition to the 12-member executive board, the development of the new doctorate will also come under the scrutiny of six scientific trustees, including Nobel laureates Renato Dulbecco and Werner Arber. □

Russian science bodies still at loggerheads

Moscow. The Russian Academy of Sciences has backed suggestions from a group of experts from the Organization of Economic Cooperation and Development (OECD) that there should be a pruning of the scientific community and a weeding out of less effective scientific institutions.

The academy's recommendation is one of a number of suggestions presented to Russian President Boris Yeltsin in response to his request for proposals from both the academy and the Ministry of Science on improving the organization of Russian science (see *Nature* 365, 283; 1993).

The two institutions presented their replies at the end of last month. Some of their

recommendations are similar. For example, both strongly criticize the way in which funding for scientists is now distributed, which the ministry in particular considers to be the main reason for the disappointment of scientists in the president's policies.

Both agencies also advise Yeltsin to stop financing ineffective scientific institutions. But the academy states its position much more strongly than the ministry. It says that a thorough review of scientific institutions is necessary, after which "ineffective research groups should be closed down".

The academy also recommends that control of all basic research in Russia should be placed under its leadership. Indeed, top acad-

emy officials have been urging Yeltsin to sign a decree confirming its authority.

In addition, the academy would also like to be given responsibility for the budgets allocated to university research, leaving to the Ministry of Science the management of the state scientific and technical programmes.

In contrast, the ministry, headed by the Minister of Science, Boris Saltykov, has disagreed with the OECD's conclusion that Russia has too many scientists (and scientific institutions), apparently worried that drastically reducing their numbers could have serious social consequences (see *Nature* 365, 379; 1993).

In contrast to the academy, with which it remains in conflict, the ministry does not make any radically new proposals. Its main suggestion is that public spending on research should increase from three to four per cent of the state budget. And the ministry also repeats its suggestion for the creation of a system of "state professors".

Previously the ministry had suggested that 5,000 such posts be created; in the current proposal, this has been increased to between 10,000 and 15,000. In order to carry out this programme, the ministry proposes to create a special Foundation of Basic Research to be given 5 per cent of the total scientific budget of Russia. But has ignored the idea of paying stipends of 75,000 rubles (about \$70 a month) to "outstanding scientists" which has been proposed by the academy and approved by Yeltsin.

The ministry has also given up the idea that academy institutes might be included in a system of federal scientific centres. It says that it plans to submit a proposal for another system of such centres — this time purely within the academy. **Vladimir Pokrovsky**

US 'should re-create national bioethics forum'

Washington. The US Office of Technology Assessment (OTA) has proposed that Congress should re-establish a national forum for debating the ethical implications of biomedical research and medical practices. The OTA is not charged with making formal recommendations to Congress but says it favours the creation of a broad-based commission over a more *ad hoc* approach, which should be considered only as a "last resort".

The OTA's analysis is made in a report, *Biomedical Ethics in US Public Policy*, details of which were released last week at a hearing held by the Senate Labor and Human Resources Committee.

It has been four years since the last government-sponsored bioethics forum, the Biomedical Ethics Advisory Committee (BEAC), went out of business. It did so after congressional members of the Biomedical Ethics Board, set up to oversee BEAC, became deadlocked along partisan pro-choice/anti-abortion lines.

Previous federal ethics bodies had some success in the 1970s and early 1980s. But however successful these bodies may have been, they did not end the need for further study, says Alexander M. Capron, of the University of Southern California Law Center, Los Angeles.

Those who support the notion of re-establishing a bioethics forum say that topics such as the conduct of AIDS vaccine trials, the compassionate uses of gene therapy, germline gene therapy, and the patenting of human tissues, cells or DNA are just a few of the areas that may raise unresolved ethical issues.

As part of its study, OTA has identified a number of factors that, irrespective of the type of body established, appeared to be crucial for the success of future efforts. Foremost among them, it says, is the need for adequate funding. OTA estimates that a standing body located within an existing

government agency could cost about US\$750,000 a year to run. One that would operate as an independent commission for a set term could cost less than US\$2 million a year for four years.

The experience with BEAC has, not surprisingly, left many in Congress reluctant to reconsider the issue of bioethics. The last attempt "really left a bad taste in a lot of people's mouths" says a congressional aide to Senator Mark Hatfield (Republican, Oregon). Nevertheless Hatfield, one of three senators to request the OTA study, introduced a bill in May that would establish a 15-member ethics advisory board within the Department of Health and Human Services. Although Hatfield says that he is not wedded to the idea of a permanent ethics advisory board, he hopes that introducing his bill will serve to focus the debate within Congress.

Diane Gershon

Minister disbands German health bureau

Munich. The German parliament is to decide whether to accept a proposal from the health minister, Horst Seehofer, to disband the federal health bureau, which he says is not fulfilling its advisory function, and require the research institutes it runs to report directly to government ministries.

Seehofer recently dismissed the president and director of the bureau following a row over HIV-infected blood (see *Nature* 365, 594; 1993). Now he says that these posts will not be filled, and that the bureau's 300-strong Berlin-based central administration, which controls more than 3,000 staff in six major institutes and several working

groups, should be dismissed.

In the shake-up Seehofer is proposing, six institutes would become directly answerable to the health minister, and one be transferred to the environment ministry. As well as conducting research, each of these institutes helps to set and control standards in different health areas.

Seehofer's row with the bureau earlier this month did not reveal any new statistics for the number of people contracting AIDS through contaminated blood or blood products. But wide press coverage renewed public alarm, with many patients electing to cancel routine operations.

In response Seehofer says he intends to set up a DM10 million (US\$4 million) fund to help victims. He has also promised an independent inquiry into all incidences of AIDS infection from blood sources, to be completed by Christmas. **Alison Abbott**

IMAGE
UNAVAILABLE
FOR
COPYRIGHT
REASONS

Horst Seehofer

Common inheritance

SIR — A leading article (*Nature* 362, 379; 1993) on the subject of the world's human population growth stated that any concerted effort by the international community to stabilize growth rates in poor, developing countries will not be successful unless accompanied by a programme of mass education and public health, and by the elimination of trade barriers. The concern about the growth of the world's human population follows from the demands that such increases impose on the 'global commons'. Although a necessary part of the solution, such measures are not sufficient. We cannot solve the problems of human population growth and its effects on the environment in isolation; we must simultaneously address the issue of energy consumption. The leading article failed to do so and one is left wondering whether the omission is the result of ignorance or ideology.

An analysis of the patterns of energy use broken down by country/geographical region was published recently in a report produced jointly by the World Conservation Union (IUCN), the United Nations Environment Programme (UNEP) and the World Wildlife Fund (WWF)¹. Although three-quarters of the world's population live in relatively poor, developing countries, they account for only 20 per cent of the world's commercial energy consumption.

The remaining 80 per cent is consumed by people living in relatively rich, high energy consumption countries, most of which continue to show increases in resource consumption despite having near stable populations¹. The per capita commercial energy consumption rate in a high consumption country is 18 times higher than in a low consumption country. On average, a North American produces twice as much carbon dioxide emission as a South American and ten times more than someone living in South Asia or East Asia (with the exception of Japan)¹.

These data clearly show that action by the international community aimed only at curbing fertility rates in developing countries will not solve the problem of excessive demands on the world's natural resources.

There is no doubt that the rapid population growth exhibited by many low energy consumption countries undermines the prospects for sustainable development. However, addressing this problem while neglecting the need to curb habits of energy use in richer countries will not lead to a reduction in the demands on the global commons. In fact, one might argue that these problems would actually increase if poorer countries were to slow their birth rates and adopt the energy use patterns of richer countries.

Both reduction in energy use and the curbing of fertility rates are extremely difficult policies to implement. However, it is in part the unfairness of the demand that only one group of countries address the problem by implementing unpopular policies that raises the hackles of people speaking on behalf of these countries. Co-operation from less developed countries in terms of implementing policies to curb the birth rate might be more likely if a real commitment to reduction in energy consumption and its consequent damage to the environment was evident in the policies of richer countries.

Daniel E. Ruzzante

Sandra J. Walde

Department of Biology,
Dalhousie University,
Halifax, Nova Scotia,
Canada B3H 4J1

1. *Caring for the Earth. A Strategy for Sustainable Living.*
IUCN/UNEP/WWF (Gland, Switzerland, 1991).

Anglo-Saxon attitudes

SIR — I have been much interested in two book reviews in *Nature*, separated by a year, which reflect what I think is an interpretation bias about what French and other continental Europeans feel about the Anglo-Saxon scientific establishment^{1,2}. Coming from a small country, my reaction cannot be ascribed to delusions about past "grandeur". Of course, we all know of papers submitted to journals, retained a long time by referees, then rejected, only for us to see their main contents appearing in the same or another journal with the signature of what we guess was the referee.

But as referees are anonymous, this is difficult to prove. The error in the *La Science au Présent*, which *Nature* recently reviewed², is to suppose that this happens to French authors because they are French. It happens to all of us, including US researchers themselves. This reflects poor ethics and the results of fierce competition and of *Science Citation Index* ideology, but not national discrimination. On the other hand, as *Nature* seems sensitive to such allegations, I wondered if it is immune to national or cultural discrimination. I therefore looked at the origin of the items which result from editors' requests and not from submissions, that is, News and Views, book reviews, commentaries and reviews. The results of a 24-month survey (from 30 May 1991 to 27 May 1993) were:

■ News and Views: UK 299, USA 349, "Commonwealth" (Canada, Australia,

New Zealand, South Africa) 10.5, Continental Europe 25, Japan 2.5.

■ Commentaries: UK 8, USA 7, Commonwealth 3, Continental Europe 1.5, Japan 1.

■ Reviews and Hypothesis: UK 12.5, USA 40.5, Commonwealth 2, Continental Europe 4.5, Japan 0.

■ Book reviews: UK 351, USA 161, Commonwealth 15, Continental Europe 12, Japan 1.

This excludes *Nature* editors. When authors of one item are of 2 nationalities, a fraction is apportioned to this nationality.

A bit biased for an international journal³, is it not? and may we congratulate the few "continental" contributors?

E. A. Dumont

University of Brussels,
Brussels,
Belgium

1. Smith, P. J. *Nature* 354, 27 (1991).

2. Cols, P. *Nature* 362, 680 (1993).

3. Maddox, J. *Nature* 359, 475 (1992).

Forest cover

SIR — Although not diminishing the importance of proposals for forest conservation in Haiti made in Correspondence by Hedges and Woods¹, there is a serious error in the figure cited for Jamaican forest cover. In 1980, a comprehensive Jamaican land cover/use assessment revealed 46.22 per cent forest cover², and although the decrease in forest cover from 1980 to 1986 was 3.3 per cent per year and has continued at around this level since³, the fraction of forested land in Jamaica is about eight times higher than the 5 per cent shown by Hedges and Woods.

Jamaica retains a larger fraction of its forest than other Greater Antillean island despite a population density equal to Haiti because of three unique geographic characteristics: Jamaica is generally higher, wetter and mostly limestone (around 80 per cent). Jamaica (the "Island of Wood and Water" is the literal translation of the Arawak Indian name *Xaymaca*) remains greenest only because lucky geological accidents cause most forest to lie on land too steep, too rainy or on limestone too lacking in soil to be worth clearing. Habitat degradation and erosion are severe on cleared Jamaican soils, but there is still a good chance to preserve what is left before it nearly vanishes like the tropical forest in Haiti.

T. J. Goreau

Global Coral Reef Alliance,
324 North Bedford Road,
Chappaqua, New York 10514, USA

1. Hedges, S. B. & Woods, C. A. *Nature* 364, 375 (1993).

2. Rampair, S. in *Forests of Jamaica* (eds Thompson, D. A., Bretting, P. K. & Humphreys, M.) 77–80 (Jamaican Society of Scientists and Technologists, Kingston, 1986).

3. Eyre, L. A. *Jamaica Nat.* 1, 27–44 (1991).

Electronic journals are already here

This year's gathering of publishers at Frankfurt has made it plain that the electronic distribution of research data is no longer in the future, but in the here and now.

Frankfurt. After several years in which publishers have been proclaiming that electronic publishing is around the corner, many now seem willing to acknowledge that it has arrived. Indeed, at this year's Book Fair, a whole exhibition has been given over to electronic publishing in various forms. Some commercial publishers are plainly prepared to get their feet wet, and to risk their capital as well, in this brave new world. But, wisely, nobody is pretending that what is now on show is what the book fair will be like a decade or so from now.

Indeed, by the standards of the research community's use of Internet and the other readily accessible electronic networks, the ventures of general publishers into the electronic future are elementary at least. What has happened is merely that several publishers have recognized the convenience of compact discs as a means of distributing data for which there is a general market, dictionaries and other reference works for example. For the time being at least, there are not many competing formats to confuse the customer (and complicate the manufacturing process). With bit-mapped images, the capital cost of piracy is high, comparable with the cost of pirating a printed book. And selling large numbers of identical compact discs is not very different from selling books. It is easy to feel technically advanced when nothing much has changed.

Whether the few signs at this year's fair of a more complicated time ahead are to be taken seriously is far from clear. Some general publishers have been captivated by the multimedia craze, offering compact discs carrying both text and illustrations. The snags, for the time being, are that the quality of the illustrations is not high and their interleaving with the text is artificial, much as if a book illustrator had been commissioned to produce 'something' every eight pages or so. But that could change. So could the sophistication of educational packages intended both as textbooks and compilations of relevant data. That is where the general publishers are likely first to strike electronic gold.

The future of scientific publishing is altogether a different question, if only because of the access to digital electronic networks which the research community has enjoyed for many years, and which it now relies upon. Although compact discs are sometimes used (especially, for example, for the transfer of large amounts of numerical data in digital form), they already have an old-fashioned air. Who wants to be

hunting for physical discs in a small filing cabinet when it should be possible to retrieve the same data directly onto a screen simply by typing a remembered filename?

Whence the general sense that scientific journals (and their publishers) are heading for a revolution. This year's symposium of the STM (for "science, technology and medicine") publishers' group, ostensibly about the "information explosion" in the field of scholarship as a whole, could not keep off the question whether true electronic networking of scientific texts will blunt whatever difficulties there may be in keeping up with the literature. One participant claimed that there are already 130 strictly electronic journals, but there is no listing of them.

The plain truth of course, is that the publication of research material by means of electronic networks is inevitable. Indeed, in some fields, it is already standard practice. In both particle physics and astronomy, it has become standard practice for people to list the titles of their latest preprints on bulletin boards, from which copies can be retrieved electronically.

Opinions of this informal service vary. Some say that they now use the journals in their field only to look up articles published before the coming of the networks, others that they prefer to judge for themselves whether a paper makes sense and do not set much store by its validation by referees before its publication as ink on paper — and still others that the information explosion is an illusion, given that they learn all they need to know from their graduate students and on the telephone — the 'personal communication' of the formal paper.

These developments raise awkward problems, both for the research community and for publishers. There is every likelihood that it is a matter of a few years only before the practice long-established in particle physics has spread to many other fields. (The frequency with which biologists now sport Internet addresses is already conspicuous.) But the question will then arise of whether the formal publication of research articles in what are at present called journals will continue to make sense.

Even in particle physics, there are many who still hold that formal publication will always be necessary. For how else are future cohorts of researchers to be guided towards the literature worth taking seriously? Certainly those who make appointments and promotions, and who award research grants, continue to set great store by publication in journals that have been rigorously refereed.

But will referees continue to perform their invaluable unpaid services with a good grace, or at all, if they know that they are commenting on research articles that have already been distributed to all those likely to be interested, and which are unlikely to be read a second time even if they are substantially changed after being refereed? That seems improbable. In short, when access to the networks is universal, it seems unlikely that the concept of the journal as an authenticating agent will survive even in an electronic form.

In the circumstances, it is perhaps as well that there remain a few obstacles to this free flow of data and its interpretation. Thus photographs used as illustrations remain a technical problem; although compression by a factor of about 50 is possible, making handling more economical, there is still no uniformly agreed format by which that can be done. But that can be only a passing difficulty.

More seriously, there is the threat that access to the networks will not indefinitely be free of charge, and that charges may be levied that are somehow related to the use made of them. Not only the Russian scientists who have learned to live by the networks would find such a development devastating. Yet it is a curious feature of present arrangements that the research community prefers that this crucial issue should not be discussed. It is as if people believe that network charges are not yet a reality because the issue has been overlooked.

The publishers' dilemma (which in this context includes the learned society publishers of important journals, many of whom derive the bulk of their income from this source) is different and old-fashioned. Who will continue to buy ink-on-paper journals when most of the contents, admittedly in unrefereed form, have already been distributed on the networks? Seeking to outwit events by publishing journals electronically is no defence, because there is no obvious way of restricting unlicensed copying. Either way, there is a prospect of continuing overhead costs and shrinking revenue.

Luckily, again, the revolution will not happen tomorrow, or even the day after. There will be some time to adjust. But it is interesting, to judge from the Frankfurt meeting, that publishers are only now facing up to what may lie ahead. It is as if they have believed that, by saying nothing, they could avoid inviting trouble. Is it possible that, by next year, they will even be talking to the scientific community?

John Maddox

Odd solid predictions

Jeffrey S. Moore

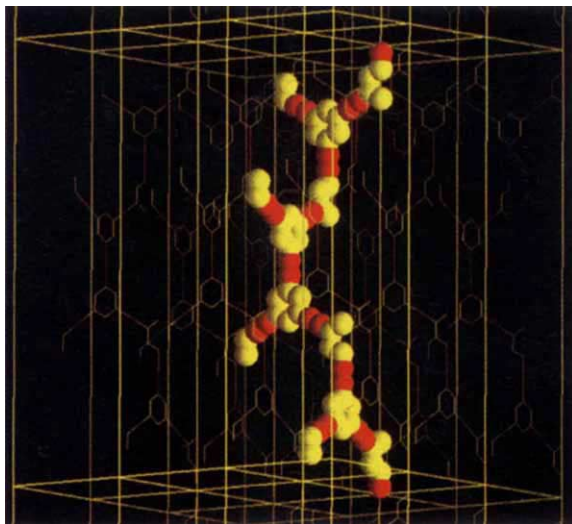
WHEN stretched, most materials become thinner in cross-section and less dense; when heated, most expand. These intuitive relationships are a consequence of the anisotropic nature of the chemical bond and the asymmetry of the potential energy surface that describes this pairwise interaction. Although there are examples of materials with anomalous thermal and mechanical behaviour, such behaviour can usually be traced to microstructural details at a length scale well above the molecular level. On page 735 of this issue¹, Baughman and Galvão describe something else again — they predict that a class of materials they call 'hinged network crystals' will have decidedly counterintuitive properties, which stem from a unique bonding connectivity.

As the name implies, hinged network crystal phases have the peculiar feature that the entire lattice can undergo reorganization by a nearly barrierless twisting type motion around a set of parallel torsions extending across the infinite lattice. What is unique about this motion is that it can occur without involving a change in the network bond lengths, so it takes place smoothly, like a hinge. It is fascinating to explore the massive scale of concerted atom displacements that pervades the entire phase with molecular models. Because the process involves purely torsional rotation, the deformation requires very little energy (it is extremely compliant). If the internal dimensions of the unit cell are large relative to the diameter of individual network strands, then the torsion requires minimal energy. That is what Baughman and Galvão call the 'uncrowded' type of hinged network phase. On the other hand, if the strands of the network fill the space of the unit cell, the torsional motion is restricted and the unusual properties may no longer be observed. In this case, the compliant deformation is constrained by repulsive van der Waals interactions and other deformation mechanisms predominate.

The structure of the proposed phases can be envisaged as arising from two sets of non-parallel chains. The two sets stack on top of one another in alternating fashion. Covalent bonding between chains in adjacent layers forms a periodic, three-connected network that extends indefinitely. The prototype network has the same structure as the known inorganic materials ThSi_2 or LaPtSi . However, these examples fall into the 'crowded'

category. An example of an 'uncrowded' hinged network crystal is shown in the figure, where the tri-connected vertices are 1,3,5-substituted arenes joined by acetylene bonds. These more open forms exhibit a soft shear deformation around the set of parallel $-\text{C}\equiv\text{C}-$ bonds that extend throughout the solid.

The mechanical and thermal properties



Connectivity diagram and space-filling model of an uncrowded hinged network crystal¹. The tri-connected vertices are aromatic rings and the edges are acetylene bonds. A molecular fragment of this network has recently been synthesized (Z. Wu and J.S.M., unpublished data; ref. 5).

predicted for these networks are unusual. Baughman and Galvão performed their calculations on a network of polydiacetylene chains which have sp^2 carbon atom vertices and $-\text{C}\equiv\text{C}-$ edges. Tensile stress along the direction of the helical axis causes the chains to untwist. As a consequence, the material exhibits a negative Poisson's ratio (in other words, it becomes thicker in cross-section when stretched)². Furthermore, this structure is predicted to become denser when stress is applied along any of the three unit cell directions and it is calculated to have a large negative thermal expansion coefficient.

The mechanisms behind these unusual characteristics are fundamentally different from those that have been invoked to account for a negative Poisson's ratio in other materials. Such behaviour has been attributed to unusual microstructure such as inverted honeycombs. And although other molecular-based networks have been proposed as giving a negative Poisson's ratio, they were essentially an extrapolation of the inverted honeycomb down to the molecular level³. The mechanism put forth by Baughman and Galvão is a completely new concept, in that it is based

on a shear deformation that results from untwisting of helical chains. Moreover, a mechanistic relationship between negative thermal expansion and a negative Poisson's ratio has previously never been postulated.

The difficulty in verifying the unique properties of these phases is that, at present, there is no way to make such phases in extended form. The rational synthesis of extended solid-state structures remains a considerable challenge. Approaches based on topochemical reactions — that is, the transformation

of a molecular crystal to extended solids — might be feasible. But special difficulties of the present phases, such as the intricate bonding connectivity and the openness (that is, non-close-packed atoms) required of the 'uncrowded' phases, make such routes unlikely. Happily, though, there is a growing determination to tackle these issues⁴, and in the short term the preparation of finite approximations to infinite networks are well within the capabilities of today's chemists. For example, a molecular fragment of the network shown in the figure has already been synthesized (Z. Wu and J.S.M., unpublished results; ref. 5).

Once these phases are synthesized, the question of applications will arise. Such materials would have special mechanical properties, including excellent toughness and resistance to damage, as well as high Young's modulus for stress along an axis that gives a negative Poisson's ratio. If the chains have extended π -conjugation, and therefore are semiconducting, it may well be that they will have interesting mechano-coupled optical and electrical properties as well.

Baughman and Galvão present an intriguing mechanism that leads to very odd mechanical properties in solids. Their approach is an example of computer-aided molecular design of materials. But to be more than just a curiosity, synthetic chemists must now extract the essential features of hinged molecular crystals and map them into real materials — which should be even more fascinating than the model. □

Jeffrey S. Moore is in the Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.

1. Baughman, R. H. & Galvão, D. S. *Nature* **365**, 735–737 (1993).

2. Lakes, R. S. *Adv. Mater.* **5**, 293–296 (1993).

3. Evans, K. E., Nkansa, M. A., Hutchinson, I. J. & Rogers, S. C. *Nature* **353**, 124 (1991).

4. Diederich, F. & Rubin, Y. *Angew. Chem. int. Edn Engl.* **31**, 1101–1124 (1992).

5. Wu, Z. & Moore, J. S. *J. Am. chem. Soc.* **114**, 8730–8732 (1992).

Vehicles for gene therapy

James M. Wilson

FOUR years ago the gene responsible for cystic fibrosis (CF) was identified, since when there has been intense research into ways of treating the disease through gene therapy. The pace of development has been unprecedented in medicine, and there are now four clinical trials underway in two continents. They involve, as delivery vehicles, either modified viruses or liposomes. The latest news comes from Zabner *et al.* in *Cell*¹, where the authors report on a trial in humans that employed recombinant adenovirus. Their results are very encouraging. But whereas both the viral and liposome approaches are promising, both have their potential limitations.

The gene defective in CF encodes a chloride channel, called the cystic fibrosis transmembrane conductance regulator (CFTR), that is expressed in epithelia of several organs. The lung has been the initial target for CF gene therapy because the pulmonary manifestations of the disease — characterized as airway obstruction, recurrent infections, and ultimately respiratory failure — are the most serious and life-limiting. Ultimately one can envisage ways of reconstituting gene expression by inhalation of the recombinant gene construct, but complexities in the biology of the diseased lung pose unique challenges. The CF gene is regulated in a highly cell-specific manner in the adult human lung, its product — the CFTR — being expressed in subpopulations of cells throughout the superficial epithelia of both conducting and respiratory airways², as well as in cells of the submucosal glands³. It will be difficult to reconstitute precisely the normal expression of the CF gene using available gene-transfer techniques, although this may not be necessary to achieve clinical efficiency. Another consideration is the stability of recombinant gene expression in the recipient — until the pulmonary progenitor cells responsible for repopulating lung epithelia are identified and made accessible for gene transfer, patients will probably require repeated treatment.

The primary goal of CF gene-therapy research in both the commercial and academic sectors is the development of a vehicle for delivering the normal CFTR gene directly into the lung. The demands on that vehicle are high. It must efficiently transfer the gene into nondividing airway epithelial cells in a manner that is safe and clinically practical. Furthermore, recombinant gene expression must be prolonged, and the vehicle must not elicit an immune response that is either destructive or precludes repeated administration.

In principle, recombinant adenoviruses have the appropriate biological properties

for such vehicles. This family of viruses, which normally causes infections of the upper respiratory tract, can be modified for gene therapy by replacing a gene that controls viral replication (*E1a* and *E1b*) with a normal CFTR minigene. Experiments in the cotton rat, an animal that resembles humans with respect to the biology of human adenoviruses, have shown that recombinant adenoviruses can indeed transfer recombinant CFTR genes into cells of the airway⁴. Human CF respiratory epithelia reconstituted *in vitro*⁵ have helped in demonstrating efficient adenovirus-mediated gene transfer and correction of the bioelectric properties of the epithelia.

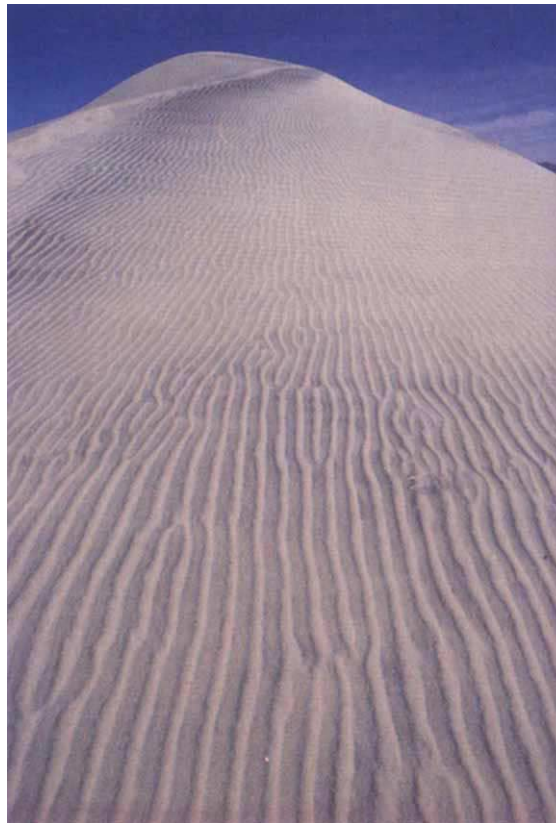
Zabner *et al.*¹ describe results of a clinical trial in which recombinant adenoviruses, containing the normal CFTR gene, were introduced into nasal epithelia of three CF patients. The rationale for evaluating the consequences of gene transfer in the nose is that nasal epithelium is an extension of the epithelium of the lung and is accessible for harvesting cells and performing direct measurements of function. The defect in chloride transport, as measured by transepithelial voltages, was corrected in all three patients.

No toxicity has been observed in another clinical trial, underway here at the University of Pennsylvania, when similar doses of adenovirus were instilled into the airways of two CF patients (J.M.W. and J. Engelhardt, unpublished data).

The observation of biological efficacy in nasal epithelia with doses of virus that do not cause toxicity in lung is heartening in view of previous studies in a variety of animal models, including non-human primates, in which there was substantial inflammation in lung with higher doses of virus⁶. But the biological effect of gene therapy in the patients of Zabner *et al.* was transient, lasting less than three weeks, which is consistent with studies in animal models. These preclinical and clinical studies suggest that therapeutic applications of the available adenoviral technology will require repeated administrations, which may lead to confounding immunological responses. An alternative strategy is to modify the recombinant virus to facilitate more stable expression of the CFTR transgene or to make the virus less immunogenic.

Liposome-mediated gene transfer has emerged as another promising way of treating CF lung disease. Several groups have used mouse models of CF generated by site-directed mutation of the CFTR gene in embryonic stem cells, to evaluate the biological efficacy of liposome-DNA complexes. Hyde *et al.*⁷ have analysed

The case of the coarsened crests



In a pattern within a pattern, the sand grains in these ripples are sorted so that the coarsest are at the crests and the finer grains in the troughs. The effect resembles the sorted stone stripes found on some hillsides (*Nature* 361, 117; 1993). In this case, though, the sand grains hop into position under the influence of the wind blowing across the ripples, whereas 'sorted' rocks are levered from their beds by the periodic growth of ice crystals. Cellular-automaton techniques applied to the stone stripes again prove successful here: grains represented by computer pixels are individually shifted according to a simple set of physical rules. Given two distinct grain sizes and a reasonable picture of how the grains leap and scatter one another on impact, R. S. Anderson and K. L. Bunas find they can pin down the cause of the sorting. And the answer? See page 740 of this issue.

LM

the effect of intratracheal instillation of liposome-DNA complexes into the CF mouse generated in Cambridge, UK, while Alton *et al.*⁸ have characterized the consequences of liposome-DNA complexes, delivered in aerosol form, in the Edinburgh CF mouse. Evidence as to the efficacy of these therapies came from measurements of the bioelectric properties of airway epithelia.

In both studies, cyclic-AMP-mediated chloride conductance, which is reduced in CF mice and humans, was at least partially corrected. Measurements of chloride transport in these studies are not easy, however, because mice, in general, have a CFTR-independent chloride conductance⁹ and in the Edinburgh mouse, in particular, there may be residual levels of functional, endogenous CFTR¹⁰. Nonetheless, an alluring feature of liposomes is that, as vehicles, they may be less toxic and immunogenic than recombinant viruses. Based in part on the results of the study by Alton *et al.*, a phase I trial of liposome-DNA complexes in CF patients has been initiated in London.

There remains a lot to be learned, and a long way to go. The main task to my mind is to pursue and solve the issues central to an objective evaluation of the effectiveness of existing gene transfer technologies for treating CF lung disease. The encouraging results obtained in human nasal epithelia with recombinant adenoviruses should be confirmed in the epithelia of the proximal and distal lung; furthermore, strategies must be developed to allow for prolonged genetic correction or repeated administration of the recombinant gene construct (or both). Finally, it is essential that the relationships between the cellular defects in ion transport and the cascade of events that leads to respiratory failure and death — bronchial obstruction, recurrent infections and dilation of airways — are better defined. Until we understand the mechanisms involved, it will be difficult to evaluate the usefulness of corrected chloride conductance as a surrogate for actual clinical efficacy. Final judgement of the usefulness of these and other gene transfer technologies must await the larger human trials that evaluate clinical endpoints such as pulmonary function. □

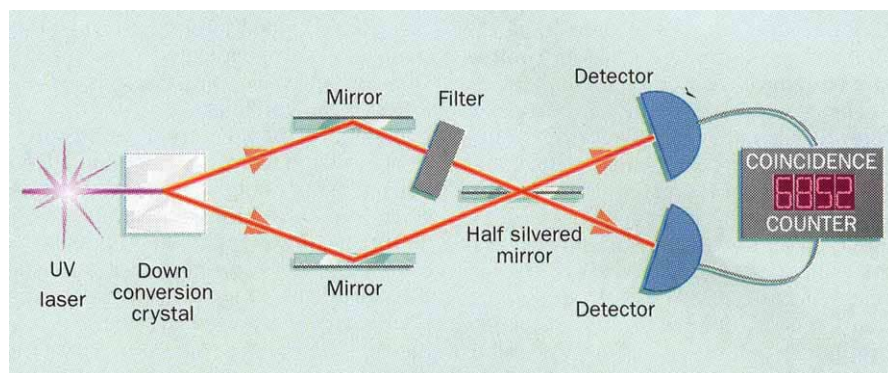
James M. Wilson is at the Institute for Human Gene Therapy, University of Pennsylvania Medical Center, Philadelphia, Pennsylvania 19104, USA.

Light faster than light?

Rolf Landauer

A ROTATING lighthouse beacon impinges on a distant wall. The resulting spot can move along the wall faster than the velocity of light. That is not a violation of relativity; one spot is not the source of the next spot. Measurements by Steinberg, Kwiat and Chiao¹ of pulse propagation through an optical filter, which reflects most of the incident light, show that the few transmitted photons arrive earlier than the velocity of light would allow, based on the simplest notions of their time

over six decades, particularly within the past decade. Many of these analyses follow a wave packet through the barrier, comparing the time of an emerging peak to the arrival time of an incident peak. It can be argued that quantum mechanics does not conserve peaks; an emerging peak is not necessarily related to the incident peak in a causative physical way. To make this convincing, a particular wave packet and barrier have been described² in which the incident peak



Two-photon apparatus for measuring relative arrival times along two paths. One path includes an optical filter which transmits only a small fraction of the incident light.

of impingement on the filter. But as in the case of the moving beacon, that velocity is presumed not to represent the retardation between a cause and its effect.

The filter consists of a periodic array of layers of material with alternately high and low refractive index. In the wavelength range from 600 to 800 nanometres, the reflections arising at the successive interfaces reinforce constructively, reflecting about 99 per cent of the incident light. The electromagnetic wave decays exponentially through the thickness of the filter.

Such evanescent waves where the decay is associated with reflection, not absorption, also occur in total internal reflection, when a wave is incident obliquely on an optically less dense medium. Evanescent waves are found in waveguides below their cut-off frequency. The reflection of most of the incident photons, and the accompanying exponential decay, are reminiscent of quantum mechanical tunnelling, when a particle is incident on a barrier that classically would cause the particle to be turned around. If the barrier is thin enough, quantum mechanics allows some chance of transmission. The analogy goes beyond that sketched here; quantum tunnelling and electromagnetic evanescent waves are described by closely related equations.

The time associated with quantum barrier penetration has been discussed for

arrives at the barrier after the transmitted peak has already left the far side of the barrier. In that particular example the small departing peak clearly arises from the front end of the incident packet.

The use of an auxiliary degree of freedom, coupled to the tunnelling process, and used as a clock, represents an alternative theoretical and experimental approach³. When incoming peaks cannot be simply physically identified with emerging peaks, then there is no reason to expect the time measured by a clock to agree with that found by following peaks.

Steinberg and co-workers¹ use an ingenious two-photon interference method for measuring the delay (see figure). Two-photon interference methods⁴ do not, as the name suggests, cause one photon to interfere with an independently generated one, but instead cause one history for the set of two photons to interfere with another history. Typically the two photons involved are generated together, and are essentially clones. If the two photons arrive simultaneously, via two different paths, at the half-silvered mirror, then the history in which both are reflected, and the history where both are transmitted, can be shown to interfere destructively. Thus, if two photons arrive simultaneously within the time resolution of the detectors, but not close enough in time to cause the destructive interference

1. Zabner, J. *et al.* *Cell* **75**, 207–216 (1993).
2. Engelhardt, J. F., Zepeda, M., Cohn, J. A., Yankaskas, J. R. & Wilson, J. M. *J. clin. Invest.* (in the press).
3. Engelhardt, J. F. *et al.* *Nature Genet.* **2**, 240 (1992).
4. Rosenfeld, M. A. *et al.* *Cell* **68**, 143–155 (1992).
5. Rich, D. P. *et al.* *Human Gene Therapy* **4**, 461 (1993).
6. Simon, R. H. *et al.* *Human Gene Therapy* (in the press).
7. Hyde, S. C. *et al.* *Nature* **362**, 250–255 (1993).
8. Alton, E. W. F. W. *et al.* *Nature Genet.* **5**, 135 (1993).
9. Clarke, L. L. *et al.* *Science* **257**, 1125–1128 (1992).
10. Wilson, J. M. & Collins, F. S. *Nature* **359**, 195–196 (1992).

just mentioned, we are more likely to detect coincident arrival. This provides a delicate tool for measuring the relative timing for the arrival of two photons. Insertion of the filter causes a change in delay, which can be measured by introducing a compensating change in the length of the alternative vacuum path. Despite the emphasis Steinberg and collaborators place on the quantum nature of this experiment, that seems to be an incidental aspect. If, instead of single photons, we had used large pulses and measured relative arrival times through nonlinear optic effects, we could have measured the same delay.

In their principal result, a delay corresponding to a velocity of $1.7c$ (where c is the speed of light) is measured. Such superluminal velocities have also been found in experiments on microwaves^{5,6}, although there the more complex geometry leads to difficulties in interpretation. The measured delay in the new work agrees with predictions for the delay of the transmitted peak, relative to the incident one.

Wave propagation is more complex than geometrical optics. We cannot easily assign a portion of the incident wave to each portion of the transmitted wave. Therefore, this experiment is not a demonstration of a real physical velocity exceeding c . We can seek an easy explanation by assuming that the whole transmitted wave comes from the front end of the much larger incident wave. Nevertheless, this experiment underlines work that remains to be done. Could the experimental details be modified so that the light pulse emerges from the evanescent region before the incident peak has even reached that region? Measurements of an interaction time of a transmitted photon with its evanescent region, via a physical clock, are needed.

But even at a fundamental theoretical level, the easy explanation suggested is unsatisfying. Are we quite sure that if we signal with photon polarization, and if we are willing to lose most of the photons in an evanescent region, that we have *no* chance of a rare superluminal signal propagation? Our understanding of this is not all that it deserves to be, although it does seem unlikely that the well accepted light velocity barrier will crumble. □

Rolf Landauer is at the IBM T. J. Watson Research Center, PO Box 218, Yorktown Heights, New York 10598, USA.

Localized learning in insects

T. S. Collett

THE insect visual system looks at the world with a spherical array of photoreceptors. Signals from receptors are fed through several layers of cells where they are transformed and combined in useful ways. But, despite lateral interactions, the spatial order of the retina is maintained — each layer of the optic lobe maps retinotopically onto the next. On page 751 of this issue¹, Dill, Wolfe and Heisenberg present an elegant demonstration that visual-shape learning in *Drosophila* follows this retinotopic mapping. They find that a rewarded shape that has been presented in one position on the retina is recognized when shown there again, but not if it is shifted to a neighbouring region of retina. The huge size difference between the insect and primate visual systems suggests that insects may lack some of the features that make our system so flexible and powerful. But because insects perform many visual tasks astonishingly well, it is instructive to discover what mechanisms insects can manage without and what are the consequences of their absence. The experiments of Dill *et al.* suggest that translational invariance is one of these missing features.

A strong case for spatially localized pattern learning in honeybees was first made by Wehner² in the 1970s, and additional evidence was later supplied by Gould³. The bees in these studies were free-flying and could decide from which vantage point and in which orientation they viewed the world, making it hard to control where on the retina a pattern would fall. Wehner coped with this problem by arranging a situation in which bees tended to hover close to a narrow feeding tube in front of the pattern. Dill *et al.* have taken a different approach, and have devised a technique for studying pattern learning in tethered flies in which the eyes are fixed in space so that visual stimuli can be positioned precisely on the retina.

A flying *Drosophila*, with its head glued to its body, was attached rigidly to a transducer which measured the torque produced when the fly attempted to turn. This torque governed the angular velocity of a cylinder surrounding the fly, so that when the fly generated torque in one direction its visual world moved across its retina in the opposite direction, as would happen in unimpeded flight. The fly could thus 'turn' towards or away from patterns fixed to the inner wall of the cylinder. Visual-pattern learning was induced by an avoidance technique: an unpleasant beam of infrared light was directed onto the abdomen. Flies could extinguish this beam by fixating the correct pattern on the inner wall, and they rapidly learnt to

discriminate between different patterns such as a triangle 30° high and a similar triangle rotated through 180°. Trained flies spent most of the time facing the rewarded pattern.

During training, the triangles were held fixed in one vertical position, 1.5°, 4.5° or 15° above or below the equator of the eye. After training, both triangles were shifted vertically by 3°, 9° or 30° to an equivalent position on the opposite side of the equator. Discrimination between the triangles was unaffected by a 3° vertical shift, an amount which is less than the vertical distance between adjacent ommatidia (4.6°). It was, however, completely disrupted by the two larger shifts.

This test dispels most doubts that some pattern learning is retinotopically localized, but it remains unclear whether all visual learning in insects is limited in this way. The triangular shapes used by Dill *et al.* are complex and are likely to be encoded by the pattern of discharge from a sizeable cluster of neurons. There might, however, be mechanisms to support translational invariance for simpler stimuli, such as oriented edges or patches of colour or unidirectional motion, which can be signalled by the discharge of one or of very few cells. In principle, this might be accomplished by learning about the output of an interneuron which is sensitive to a specific pattern over a large receptive field. Thanks to Dill *et al.*, such questions can now be answered.

The primate's ability to recognize complex patterns that are shifted over the retina is evident in the behaviour of single neurons in the inferior temporal cortex. These 'elaborate cells'⁴ respond selectively to complex shapes across a significant portion of their large receptive field^{4,5}. This property is probably not inbuilt; more likely the wiring is set up during previous experience of positional shifts⁶. That the same applies to insects is implausible. Nonetheless the study of Dill *et al.* does not fully exclude the possibility that a mechanism for translational invariance exists in insects, but experience is required to establish it.

A visual system without shift invariance needs special strategies for looking at the world. If visual memories in insects are

1. Steinberg, A., Kwiat, P. G. & Chiao, R. *Phys. Rev. Lett.* **71**, 708–711 (1993).
2. Landauer, R. & Martin, Th. *Solid State Commun.* **84**, 115–117 (1992).
3. Landauer, R. *Nature* **341**, 567–568 (1989).
4. Greenberger, D. M., Horne, M. A. & Zeilinger, A. *Phys. Today* **46**, 22–29 (August 1993).
5. Enders, A. & Nimtz, G. *Phys. Rev.* **B47**, 9605 (1992).
6. Ranfagni, A., Fabeni, P., Pazzi, G. P. & Mugnai, D. *Phys. Rev.* **E48**, 1453–1460 (1993).

1. Dill, M., Wolfe, R. & Heisenberg, M. *Nature* **365**, 751–753 (1993).
2. Wehner, R. *J. comp. Physiol.* **77**, 256–277 (1972).
3. Gould, J. L. *Science* **227**, 1492–1494 (1985).
4. Tanaka, K., Saito, H., Fukada, Y. & Moriya, M. *J. Neurophysiol.* **66**, 170–189 (1991).
5. Gross, C. G. *Phil. Trans. R. Soc.* **B335**, 3–10 (1992).
6. Foldiák, P. *Neural Computation* **3**, 194–200 (1991).
7. Zeil, J. *J. comp. Physiol.* **A172**, 189–205 (1993).
8. Collett, T. S. & Lehrer, M. *Proc. R. Soc.* **B252**, 129–134 (1993).
9. Zeil, J. *J. comp. Physiol.* **A172**, 207–222 (1993).

indeed retinotopically organized, then they are best stored and retrieved from the same vantage point. There are indications that this is so. While learning the visual surroundings of their nest or feeding site, bees and wasps perform stereotyped flights that enable views to be stored from a limited number of orientations. Returning insects searching for a remembered place take up similar orientations, so

making it easier to retrieve these views⁷⁻⁹. By restricting the number of vantage points, insects can possibly dispense with size and rotational invariance as well. It will be interesting to see whether *Drosophila* lacks these constancies too. □

T. S. Collett is at the Sussex Centre for Neuroscience, University of Sussex, Brighton BN1 9QG, UK.

MICROBIOLOGY

Some like it hot (and oily)

John Parkes and James Maxwell

PRODUCTION of hydrogen sulphide on oil platforms and in oil reservoirs can be quite a headache, not least in causing corrosion of iron and steel alloys, and reducing oil recovery rates and quality¹. There is plenty of indirect evidence that anaerobic bacteria are among the culprits. However, not only have conditions in many reservoirs been considered too harsh for bacteria to survive (with temperatures sometimes in excess of 100 °C and pressures of 450 bar or more²) but crude oil has been thought to be resistant to degradation by anaerobic microbes¹.

In their report on page 743 of this issue³, Stetter and colleagues provide fresh insight into these and other issues. They demonstrate the presence of high concentrations of hyperthermophilic bacteria in production fluids (oil-water-gas emulsions) emerging from four separate oil reservoirs; the reservoirs concerned lie 3,000 metres below the bed of the North Sea and below the permafrost surface of the North Slope of Alaska.

There has been debate about reservoir microbiology since Zobell⁴, in the 1930s, reported the presence of sulphate-reducing bacteria in oil strata, and the effects of in-reservoir biodegradation on oil composition were recognized in the late 1960s (ref. 5). Sulphate and sulphur-reducing bacteria are anaerobes and obtain energy from the oxidation of organic acids, hydrogen and other low-molecular-weight compounds, coupled with the reduction of sulphate to hydrogen sulphide⁶. So they are likely candidates as a cause of sulphide production within oil fields.

There are two main ways for bacteria to enter oil reservoirs. First, they could be introduced during drilling, or during the injection of sea water used to stimulate oil production. This was, however, thought to be unlikely for hot (more than 100 °C) reservoirs where the temperature is above the limit for sulphate-reducing bacterial

isolates⁷, unless the temperature dropped below 45 °C (ref. 8) as a result of water injection. Second, they could be indigenous and have either been deposited with the original sediment and survived over geological time⁹, or have migrated into the reservoir through faults and oil seeps. Such bacteria may have been dormant or almost inactive, and have then

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Crude connection — are anaerobic bacteria a cause of sulphide production in oil fields?

become stimulated by changes in reservoir conditions caused by oil recovery, such as the introduction of sulphate and other nutrients by injection of sea water. In addition, there is the possibility of samples being contaminated by bacteria growing in the cooler parts of the production system.

Stetter *et al.* provide a new slant on these old problems in that they show that hyperthermophiles, including sulphate- and sulphur-reducing bacteria, emerge with the production fluids, and that some of these bacteria are similar to archaeobacteria found in other hydrothermal environments¹⁰. Although the existence of these bacteria, some of which are new species, does not challenge the currently accepted upper temperature limit for life of about 110 °C (ref. 10), some are clearly able to grow in media under reservoir conditions (at least 90 °C and 420 bar); and, as anaerobic sulphate reduction has been demonstrated to occur at 105 °C in another high-temperature environment, this process may occur even in these hot

reservoirs. The high temperature requirement of the isolates also means that contamination is an unlikely explanation for the results.

Stetter *et al.* consider that these hyperthermophiles were probably introduced into the reservoir with injected sea water. This contrasts with the conclusions of a study on thermophilic (43–78 °C), spore-forming, sulphate-reducing bacteria in North Sea oil production waters⁹, which were thought to be indigenous (partly because of previous reports of the absence of thermophiles in North Sea water⁸). Stetter *et al.* have, however, some evidence for the presence of hyperthermophiles in injection water. The survival of hyperthermophiles in cold (6 °C) water may seem surprising, given their average minimum growth temperature of 60 °C (ref. 3), but such bacteria are able to survive at low temperatures for years, and some anaerobes can even tolerate oxygen more effectively under these conditions¹⁰. So transport from one high-temperature environment to another in cold sea water does appear to be possible.

A twist, however, arising from the work of Stetter *et al.*, is that given the significant output of hyperthermophilic bacterial cells from reservoir production waters such cells may be able to inoculate new oil reservoirs and other high-temperature marine environments, such as mid-ocean vents and volcanoes. This export of bacterial cells indicates considerable productivity within the reservoir system, and raises doubts as to whether the concentrations of volatile fatty acids and hydrogen present there would be sufficient to maintain this activity. Some of the enrichments obtained by Stetter *et al.* appear, however, to be able to grow anaerobically on crude oil, thus apparently extending reports of anaerobic degradation of up to C₂₀ alkanes by mesophilic sulphate-reducing bacteria¹¹. But even if anaerobic bacterial degradation of crude oil is confirmed, there must surely be limits to such

Science Photo Library

1. Herbert, B. N. in *Microbial Problems in the Offshore Oil Industry* (eds Hill, E. C., Shennan, J. L. & Watkinson, R. J.) 63–71 (Wiley, Chichester, 1987).
2. Shennan, J. L. & Vance, I. in *Microbial Problems in the Offshore Oil Industry* (eds Hill, E. C., Shennan, J. L. & Watkinson, R. J.) 73–91 (Wiley, Chichester, 1987).
3. Stetter, K. O. *et al.* *Nature* **365**, 743–745 (1993).
4. Zobell, C. E. *Producer's Mon. Penn. Oil Prod. Ass.* **22**, 12–29 (1958).
5. Winters, J. C. & Williams, J. A. *J. Am. chem. Soc. Div. Petrol. Chem.* **14**, E22–E31 (1969).
6. Widdel, F. & Hansen, T. A. in *The Prokaryotes* Vol. 1, 2nd Edn (eds Balows, H. G. *et al.*) 583–624 (Springer, Berlin, 1992).
7. Jørgensen, B. B. *et al.* *Science* **258**, 1756–1757 (1992).
8. Stott, J. F. D. & Herbert, B. N. *J. appl. Bact.* **60**, 57–66 (1981).
9. Rosnes, J. T. *et al.* *Appl. envir. Microbiol.* **57**, 2302–2307 (1991).
10. Stetter, K. O. *et al.* *FEMS Microbiol. Rev.* **75**, 117–124 (1990).
11. Aeckersberg, F. *et al.* *Arch. Mikrobiol.* **156**, 56–61 (1991).
12. Erlich, H. L. & Ghiorse, W. C. (eds) *Geomicrobiology* **7** (1989).
13. Cragg, B. A. *et al.* *Proc. ODP, Sci. Res.* **127/128**, 761–776 (1992).
14. Gold, T. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6045–6049 (1992).

activity under natural conditions or all oil in reservoirs would be extensively biodegraded.

The new observations³ do not completely rule out the presence of indigenous bacterial populations, and it would be interesting to know whether there is a correlation between the start and duration of seawater injection and increases in numbers of bacteria emerging with production fluids. Also it would be important to be able to confirm the activity of bacteria under *in situ* oil reservoir conditions, and to determine their contribution to sulphide production compared to that of geochemical processes¹. The results

presented do, however, provide firm evidence for the presence of a subterranean biosphere in oil reservoirs; moreover they are consistent with demonstrations of the existence of other deep biospheres in aquifers¹² and marine sediments¹³, which together indicate that the biosphere is not just a thin veneer on the geosphere. They are bound to fuel the debate on the existence of a hot, subterranean biosphere¹⁴. □

John Parkes is in the Department of Geology, and James Maxwell in the Department of Chemistry, University of Bristol, Bristol BS8 1RJ, UK.

LIGHT-EMITTING DIODES

The silicon chameleon

Leigh Canham

GOOD news for those working on applications of light-emitting diodes for complex displays: the colour of light emitted by a porous silicon diode can vary appreciably depending on the applied voltage, say A. Bsiey *et al.*¹.

Porous silicon has excited widespread interest since 1990, when its striking light-emitting properties were first reported and ascribed to quantum size effects². It is generally created by using hydrofluoric acid to etch an array of holes ('pores') electrochemically into crystalline silicon wafers. Electron microscopy soon revealed the resulting columnar microstructure³ of luminescent material. The skeleton left behind retains the substrate crystallinity and can have very low dimensionality, as the pores typically have widths of 2–20 nanometres and are closely packed together. Structures of high porosity (void fraction) are capable of delivering very efficient (1–10 per cent) visible photoluminescence; this is on a par with well-established luminescent semiconductors such as GaAs. For applications involving light-emitting diodes (LEDs), however, efficient electroluminescence has to be achieved. In 1992, this was shown by two independent groups^{4,5} to be feasible, and the LED structure used was the one studied by Bsiey and colleagues. It is remarkable for a number of reasons.

To start with, at first sight it is not clear why it should work at all. In this device, a liquid electrolytic contact containing a specific redox species (the persulphate $S_2O_8^{2-}$ ion) infiltrates the porous layer on the bulk silicon substrate. When a cathodic (negative) voltage is applied to the silicon with respect to a metallic anode (platinum or other metal), visible electroluminescence is obtained from layers that produce visible photoluminescence. But such porous structures in air are known to be highly resistive. So when we fill the

pores with a conductive liquid, why doesn't virtually all current flow through that liquid into the bulk silicon, effectively short-circuiting the resistive porous silicon? Indeed, that is precisely what happens when the layer is being formed in the etchant.

Second, as Bsiey and colleagues show, a single layer can produce any of four colours depending on excitation conditions. The range of spectral tunability is much greater than seen previously⁵. For their layers, fabricated under illumination, a shift of electroluminescence as large as 270 nm is obtained for an external bias variation of only 0.6 V. This compares extremely favourably with tunable colour devices using stacked zinc sulphide phosphor structures without filters, where red-green colour variation requires a 30-V change in bias⁶. Third, it is remarkable that the electroluminescence output band can be spectrally so much narrower (its full width at half maximum is about 0.25 eV) than the photoluminescent band of the same layer (about 0.6 eV).

The first quandary is explicable if the electrolyte does two things. Not only do its persulphate ions efficiently inject holes into the silicon skeleton, but under cathodic bias its cations (such as Na^+ and H^+) dope that skeleton and thereby dramatically lower its resistivity. They do this by charging the surfaces of the silicon columns. In much the same way that electrolytes can induce very high electron accumulation at the surface of bulk silicon⁷, a one-dimensional electron gas is generated in the interconnected columns of porous silicon. Efficient electroluminescence at low biases can then result from recombination of localized holes ('minority carriers') with delocalized electrons ('majority carriers') that flow along the wires.

The second puzzle can be addressed within the framework of quantum-

confinement effects. A key prediction of such models is that light emission wavelength is a function of size. The narrower the 'quantum wire' or 'quantum dot', the larger the bandgap and the shorter the wavelength of emission. Theoretical calculations suggest that a 30-Å bare silicon quantum wire should emit deep red light. Native oxide growth can reduce the silicon core to about 20 Å and result in orange emission. Even narrower structures would then correspond to yellow and green emission, the band gap being very sensitive to wire size and shape in this regime. Porous silicon layers are inhomogeneous in the sense that they generally contain a range of wire widths ('crystallite sizes'). Bsiey *et al.*¹ suggest that this diode structure, where contact is made to the whole of the nanostructure, is capable of selective excitation — spectral tunability is controlled by the cathodic bias which enables smaller and smaller crystallites to be excited. Hence the dramatic blue-shift with increasing bias.

That brings us to the third puzzle. One important issue not addressed by Bsiey *et al.* is the role of vertical inhomogeneity in the layer. Many luminescent structures show strong evidence for porosity gradients with depth⁸. This is a consequence of their formation mechanism: the top of the layer, having been in the etching solution longer than the bottom, has slightly wider pores. The silicon skeleton is then smaller at the top of the layer, resulting in shorter-wavelength emission. If this effect is important, the red emission could predominantly arise from the bottom of their layer, the green from the top. Then the increasing cathodic bias, by rendering more and more of the layer conductive, would generate a blueshift. How sensitive, I wonder, was the photoluminescence to excitation wavelength and intensity?

Scientists striving to make use of semiconductor nanostructures in novel devices generally apply lithographic techniques. Beautifully regular arrays of structures with narrow size distributions are usually used as a yardstick of the success of the fabrication process. So it is ironic that here we have a nanostructure in which, as Bsiey *et al.*¹ suggest, the very inhomogeneity produces the desirable characteristic of tunable colour emission. □

Leigh Canham is at the Defence Research Agency, St Andrews Road, Great Malvern WR14 3PS, UK.

1. Bsiey, A. *et al.* *Phys. Rev. Lett.* **71**, 637–640 (1993).
2. Canham, L. T. *Appl. Phys. Lett.* **57**, 1046 (1990).
3. Cullis, A. G. & Canham, L. T. *Nature* **353**, 335–338 (1991).
4. Bressers, P. M. M. C. *et al.* *Appl. Phys. Lett.* **61**, 108 (1992).
5. Canham, L. T. *et al.* *Appl. Phys. Lett.* **61**, 2563 (1992).
6. Fukao, R. *et al.* *IEEE Trans. Electr. Dev.* **40**, 64 (1993).
7. Tardella, A. & Chazalviel, J. N. *Phys. Rev.* **B32**, 2439 (1985).
8. Inoue, K. *et al.* *Jpn Appl. Phys.* **31**, L997 (1992).

RÉSUMÉ

Joint proposal

UNUSUALLY high joint flexibility (hypermobility) can be a good thing and a bad thing for a performing musician, depending on which joints are especially flexible and which instrument is played. L.-G. Larsson *et al.* (*New Engl. J. Med.* **329**, 1079–1082; 1993) have analysed data on 660 staff and students at a school of music in New York state. The authors find that hypermobility of joints undergoing repetitive motion is an asset in reducing adverse effects in those joints. For instance, it may well benefit violinists to have highly flexible joints in the left hand in particular (as was said to be the case for Paganini). The downside is that over-flexibility in areas that provide support, such as the spine or knees, is associated with a higher incidence of musculoskeletal problems in such places. Violent conductors may pose a more serious risk of injury to a career musician — but it helps to have the right joints.

Light fantastic

FULLERENES from sunlight? It sounds like a physicist's pipe-dream, but two reports in the *Journal of Physical Chemistry* (**97**, 8696–8700 and 8701–8702; 1993) suggest that direct evaporation of a graphite target by concentrated sunlight will be a commercially useful way of generating large amounts of the smaller fullerenes. L. P. F. Chibante and colleagues use a small parabolic mirror on a tracking mount, whereas C. L. Fields and collaborators employ a 10-kW solar furnace. In both cases, argon flowing past the target sweeps the vaporized carbon away to condense in a darker region. The advantage of this, say Chibante *et al.*, is that it avoids the photochemical destruction of newborn fullerenes that is the bugbear of carbon-arc generators.

Back transmission

BOVINE spongiform encephalopathy (BSE) is thought to have been initially transmitted to cows through feed contaminated with scrapie, a sheep disease. J. D. Foster *et al.* now present the first report of experimental transmission of BSE back to the supposed source (*Vet. Rec.* **133**, 339–341; 1993). Foster and co-workers used 'positive' and 'negative' lines of Cheviot sheep, so named for their genetic differences in susceptibility to the SSBP/1 strain of scrapie; Anglo-Nubian goats were also included in the experiments because of their uniform response to several different scrapie isolates. The results with the goats were as expected, and consistent with other studies. Surprisingly, however, both lines of sheep developed the disease with an unpredictable and sporadic incidence, meaning that the link between genetics and susceptibility may well be more complex than previously indicated.

MOLECULAR MOTORS

One giant step for kinesin

Jonathon Howard

ON page 721 of this issue¹, Svoboda *et al.* report the first glimpse of a biological engine turning over: by recording the movement of a single kinesin molecule with extraordinary precision, they have directly observed the stepwise motion of this motor protein along the surface of a microtubule. An 8-nanometre step is small compared to the length of a microtubule (up to 100,000 nm). But it is a giant step for kinesin, the motor domain of

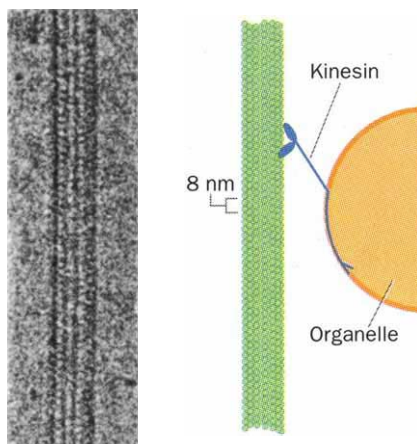
electrical work, and G-proteins which use the energy of GTP to do informational work by controlling the timing of signalling reactions. These proteins are examples of the molecular machines that distinguish cells from inanimate objects, and there is great interest from biologists, physicists and chemists in understanding the way in which they work.

Up to now, the elementary motor reaction has been obscured by the sheer numbers of motor molecules in active tissues and cells such as muscle and sperm. Svoboda *et al.* have now been able to see this reaction — the hydrolysis of a single molecule of ATP by a single motor protein — by using purified proteins in a cell-free system. By combining a laser interferometer with optical tweezers, they have made an instrument that can detect displacement with nanometre precision and millisecond temporal resolution. This feat ranks with that of Neher and Sakmann in developing the patch-clamp technique for recording the current flowing through single ion-channel molecules². Indeed, in many ways the mechanical recordings are even more difficult: in the 10 milliseconds or so during which a motor protein hydrolyses a molecule of ATP, the flow of ions through an open sodium channel will have dissipated an energy equivalent to the hydrolysis of some 10,000 molecules of ATP.

The irregular, stepwise motion of kinesin along the microtubule is evident from the time traces recorded by Svoboda *et al.* at low ATP concentration or at high force (the statistical analysis of the noisy records obtained at high ATP concentration and low force is less convincing). The spatial periodicity of about 8 nm is consistent with kinesin taking steps of 8 nm, a distance which makes a lot of sense from a structural and biochemical point of view.

The microtubule is a polymer of the heterodimeric protein tubulin. The dimers are arranged head-to-tail to form a protofilament; and the lateral association of typically 13 protofilaments forms a sheet, closure of which defines the cylinder that is the microtubule. Kinesin moves parallel to the long, protofilament axis of the 13-protofilament microtubule. The distance between consecutive kinesin-binding sites is therefore either 4 nm (the intermonomer spacing) or 8 nm (the interdimer spacing)³. The 8-nm steps are consistent with biochemical studies which show just one kinesin-binding site per dimer⁴.

As ATP concentration is decreased, the time interval between each consecutive step gets longer. So it seems that there is at most one step for each molecule of ATP



a, Electron cryo-micrograph of an unfixed 13-protofilament microtubule in ice (courtesy of R. H. Wade). The vertical striations correspond to the protofilaments. The individual tubulin monomers that make up each protofilament are just visible. b, Model of the microtubule in which the monomers are depicted as spheres of diameter 4 nm. The 8-nm steps made by a two-headed kinesin molecule as it carries an organelle along the microtubule correspond to the spacing of the tubulin dimer.

which is only 10 nm across, and it is a huge step in our understanding of how biological motors work.

Kinesin proteins such as myosin, dynein and kinesin are enzymes that convert the chemical energy derived from hydrolysis of the γ -phosphate bond of ATP into mechanical work used to power cellular motility. Myosin drives muscle contraction; dynein propels the beating of sperm and cilia; and kinesin transports organelles from one part of the cell to another along microtubules. The motor reaction is far more complex than simple catalysis, traditionally thought of as the job of proteins. Rather than just accelerating an already energetically favourable reaction, the motor protein is a mechanical device which undergoes a sequence of conformational changes coupled to the sequential chemical steps in the hydrolysis of a molecule of ATP. Energy-transducing reactions are also performed by ion pumps which use the energy of ATP to do

hydrolysed, though a careful statistical analysis still needs to be done. This supports the fundamental hypothesis underlying the standard models⁵: that there is a one-to-one association between the mechanical states and the biochemical states.

This hypothesis has been challenged by Yanagida and colleagues⁶, who have noted an apparent discrepancy between the speed of movement and the rate of ATP hydrolysis. For example, a single kinesin molecule moves at a rate of approximately 600 nm s⁻¹ (at saturating ATP concentration at room temperature) yet the maximum microtubule-activated ATPase in solution is only about 3 per second per native molecule⁷. On the face of it, this would imply that each ATP causes a movement of 200 nm, some 25 8-nm steps, quite inconsistent with the results of Svoboda and co-workers. However, there is an elegant resolution of the discrepancy. Hackney and colleagues have proposed that in solution the kinesin molecule is folded up on itself such that the tail binds to and inhibits the motor domains⁸. The binding of the tail to organelles or to the glass surfaces used in motility assays *in vitro* is presumed to open up the structure and enable the kinesin to hydrolyse ATP: the cargo turns the motor on. Consistent with this idea, proteolytic removal of the tail increases the microtubule-activated ATPase to 40–80 per second⁹ (corrected to room temperature).

These are heady times for everyone involved in the study of motor proteins. High-precision recordings such as these for kinesin and others being developed for myosin¹⁰ promise to give us a detailed kinetic and mechanical description of the reaction pathway. A recently solved structure of myosin¹¹ portends the atomic resolution of the massive conformational change — up to several nanometres — needed to explain the huge motions that these motors undergo. The combination of structural studies, genetic engineering and molecular physiology should ensure that biological motors make great strides in the next several years. □

Jonathon Howard is in the Department of Physiology and Biophysics, SJ-40, University of Washington, Seattle, Washington 98195, USA.

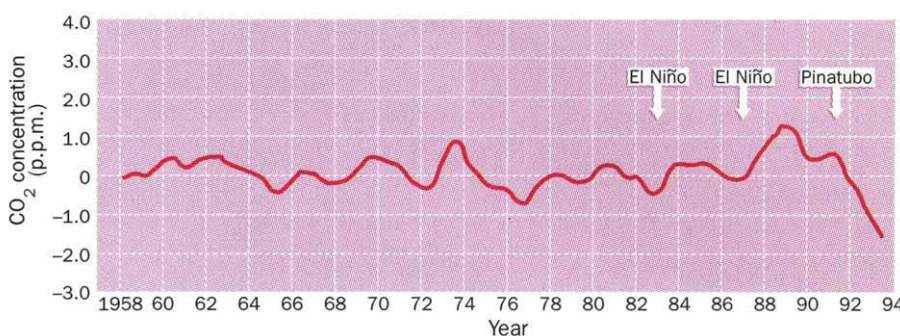
Atmospheric CO₂ stalled

Jorge L. Sarmiento

DURING the past year the growth rate of atmospheric carbon dioxide resulting from anthropogenic emissions has slowed by an amount that is unprecedented in the 35 years of David Keeling's time series of samples from Mauna Loa. Keeling unveiled these dramatic results at a meeting in Snowmass this summer*. Once the long-term trend and seasonal signal have been removed from the Mauna Loa record, the variability is strongly correlated

1991–92 El Niño should have led to increased growth of CO₂.

We can be relatively confident that the anomaly is not due to a reduction of CO₂ released to the atmosphere from fossil sources by industrial processes. Fossil-fuel emissions in 1991, estimated from the most recent United Nations statistics, were 6.19 Pg carbon, comparable to the 6.10 Pg emitted in 1990, and well within the trend of previous years (Greg Mar-



The Mauna Loa carbon dioxide anomaly obtained by removing the seasonal signal and detrending the remaining signal using a constant airborne fraction (58.58%) of the industrial release (D. Keeling, personal communication). After the two recent El Niño events of 1982–83 and 1986–87 (labelled in December of the first year), atmospheric carbon dioxide increased; but although it started to rise with the most recent El Niño, which began in 1991, there was an unexpected change in trend just after Mount Pinatubo's eruption. The slowing of the growth rate over the past year is unprecedented on this record.

with the huge changes in ocean and atmosphere circulation of the El Niño–Southern Oscillation (see figure)¹. When an El Niño event occurs, atmospheric CO₂ usually increases above its expected level. This has been attributed by Keeling to reduced uptake by terrestrial vegetation resulting from the collapse of the Southeast Asian monsoon. The ocean actually reduces CO₂ from its expected level during an El Niño because of the collapse of equatorial Pacific upwelling of carbon-rich waters. However, the terrestrial effect of opposite sign is generally larger.

The most recent El Niño began during 1991 and continued into 1992, so its effect may explain the increase in atmospheric CO₂ that began in early 1991. This time, however, instead of continuing to grow, the CO₂ began to drop in mid-1991, just after the time of the eruption of Mount Pinatubo (14–15 June). This 'Pinatubo carbon anomaly' reached –1.5 p.p.m. in Mauna Loa in May of this year; if this applied to the whole Northern Hemisphere it would correspond roughly to a loss of 1.6 Pg (1.6 × 10¹⁵ g) of carbon. The magnitude of the effect is even more remarkable when one considers that the

land, Oak Ridge National Laboratory). There is no reason to believe that the past year-and-a-half were unusual.

The cause must therefore be processes in the terrestrial biosphere or ocean. Here is where measurements of the ¹³C/¹²C ratio can help. Photosynthesis preferentially takes up the lighter isotope, increasing the atmospheric ¹³C/¹²C ratio. Oceanic uptake by gas exchange, on the other hand, has little effect on the ratio. This is true even if the reason for the oceanic uptake is biological removal of carbon in the surface ocean, because the large amount of carbon in the ocean strongly dilutes the isotopic effect of this photosynthesis. One can make use of the difference between terrestrial biosphere and oceanic behaviour to determine the relative contribution of these two processes through what David Keeling calls a 'double-deconvolution' of CO₂ and ¹³C/¹²C measurements¹.

Initially it seemed that the data on the carbon isotope ratio suggested a large oceanic sink and small terrestrial source (or reduced sink) for the Pinatubo carbon anomaly, but later corrections to the ¹³C calibration reported by Keeling at a meeting in Carqueiranne[†] now point instead to the terrestrial biosphere being more responsible.

A further source of information is

1. Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. *Nature* **365**, 721–727 (1993).
2. Neher, E. & Sakmann, B. *Nature* **260**, 799–802 (1976).
3. Ray, S. et al. *J. Cell Biol.* **121**, 1083–1093 (1993).
4. Harrison, B. C. et al. *Nature* **362**, 73–75 (1993).
5. Bagshaw, C. R. *Muscle Contraction*, 2nd edn (Chapman & Hall, London, 1993).
6. Harada, Y., Sakurada, K., Aoki, T., Thomas, D. D. & Yanagida, T. *J. molec. Biol.* **216**, 49–68 (1990).
7. Hackney, D. D., Levitt, J. D. & Wagner, D. D. *Biochim. biophys. Res. Commun.* **174**, 810–815 (1991).
8. Hackney, D. D., Levitt, J. D. & Suhan, J. *J. biol. Chem.* **267**, 8696–8701 (1992).
9. Kuznetsov, S. A. et al. *J. biol. Chem.* **264**, 589–595 (1989).
10. Warrick, H. M. et al. *Meth. Cell Biol.* **39**, 1–21 (1993).
11. Rayment, I. et al. *Science* **261**, 50–58 (1993).

*Global Change Institute on the Carbon Cycle, Snowmass, Colorado, 18–30 July 1993.

†Fourth International CO₂ Conference, Carqueiranne, France, 13–17 September 1993.

atmospheric oxygen, measured by David Keeling's son Ralph. Keeling *files* showed preliminary results and model simulations at the Snowmass meeting based on his record of atmospheric oxygen initiated at La Jolla in 1989 (ref. 2). The idea is that the ratio of atmospheric O_2 to CO_2 for a given anomaly depends on whether the anomaly is caused by changes on land or in the ocean. Synthesis of organic matter releases oxygen, and remineralization or burning consumes it in a ratio of about 1:1 on land but about 1.4:1 in the ocean. Moreover, in the oceans, dissolved CO_2 can be taken up into the large inorganic carbon pool through a series of chemical reactions, whereas dissolved oxygen is chemically neutral.

The upshot of this is that an oceanic sink for the Pinatubo anomaly would give an oxygen/carbon dioxide ratio of somewhere between 2 and 8, whereas a terrestrial sink would give a ratio of about one. In fact, Ralph Keeling's La Jolla observations show a positive Pinatubo oxygen anomaly about twice that of the negative carbon anomaly. So although it is possible that a terrestrial sink is responsible for part of the anomaly, the oxygen data imply that a significant proportion is due to the ocean.

Now we come to the difficult part: what mechanisms could cause the Pinatubo carbon and oxygen anomalies? Answering this question will first require explaining the disagreement between the carbon isotope and oxygen interpretation, but already plenty of speculations have been offered. The El Niño anomalies themselves have yet to be explained satisfactorily. For example, model estimates of the oceanic impact of the 1982–83 El Niño give a carbon uptake anomaly of only 0.8 Pg of carbon^{3,4} as contrasted with an uptake anomaly of 4.5 Pg estimated from the double-deconvolution method¹. It will be difficult to understand the Pinatubo anomaly without first understanding the background El Niño signal.

At Carqueiranne, Ralph Keeling showed that the anomaly seems to be primarily a Northern Hemisphere effect (Alert at 82° N is similar to La Jolla, but Cape Grim data at 41° S show no anomaly). His results add a further constraint: he estimates that the Northern Hemisphere oxygen anomaly between May 1991 and the end of 1992 requires a source of about 3×10^{14} moles of oxygen to the Northern Hemisphere atmosphere (or a corresponding sink in the Southern Hemisphere). This source is equivalent to a sink of 2.5 Pg of carbon in this time, or 1.5 Pg of carbon a year, if all the oxygen is released by an increase in oceanic photosynthesis. Global oceanic new production today is about 10 Pg of carbon a year.

It is unfortunate that we do not have an ocean colour imager in space at present, as a signal of 1.5 Pg of carbon a year concen-

trated in the Northern Hemisphere would almost certainly have been detectable. It would close the problem nicely if an increase in oceanic biological production could also explain the negative Pinatubo carbon anomaly, but there are two problems: first, only a small fraction of the CO_2 taken up by an increase in oceanic photosynthesis would come from the atmosphere; second, the carbon-13 data seem to suggest that the carbon anomaly is mostly terrestrial.

The coincidence of the onset of the carbon and oxygen anomalies with the eruption of Pinatubo suggests a causal connection, and one way would be iron fertilization of the oceans. Removal of 2.5 Pg of carbon would require about 0.1×10^6 tonnes of iron. Pinatubo ejected about 3 to 5 km³ of rock equivalent in the form of tephra and pyroclastics⁵; for a density of 2.8 g cm⁻³ and iron content of about 5 weight per cent, the total iron content is about 500×10^6 tonnes. So only 1/5,000 of the iron would have to be released to organisms in areas of excess surface nutrients (such as nitrates and phosphates) to provide the requisite amount. Another mechanism for a Pinatubo connection is the reduction in sunlight and the cooling caused by the eruption. The reduced radiative forcing due to reflection of light by volcanic aerosols peaks at about 4.5 W m⁻² (ref. 6) compared to the mean supply of 237 W m⁻².

Cooling accompanied by changes in rainfall could perhaps affect the balance of respiration (a CO_2 source) and photosynthesis (a CO_2 sink) on land. It could also affect the ocean's capacity to act as a carbon sink, as cooler water takes up more CO_2 . But Keeling *père* estimated that this mechanism accounts for only about 0.1 Pg of carbon.

If the anomaly is indeed attributable to the climate effect of Pinatubo, the story becomes still more interesting. We understand so little of the likely response of the carbon cycle to climate change that even discovering the effect of cooling could lead to important insights into the complex feedbacks that will surely be associated with future global warming. □

Jorge L. Sarmiento is at the Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey 08540, USA.

1. Keeling, C. D. et al. in *Aspects of Climate Variability in the Pacific and the Western Americas* (ed. Peterson, D. H.) 165–236 (Geophys. Monogr. **55**, Am. Geophys. Union, Washington DC, 1989).
2. Keeling, R. F. & Shertz, S. R. *Nature* **358**, 723–727 (1992).
3. Volk, T. *Global biogeochem. Cycles* **3**, 267–279 (1989).
4. Siegenthaler, U. & Wenk, T. in *Ext. Abstr. 3rd Int. CO₂ Conf. Environmental Pollution Monitoring and Research Programme No. 59*, 189–194 (World Meteorological Organization, Geneva, 1989).
5. Scott, W. E., Hoblitt, R. P., Daligdig, J. A., Besana, G. & Tubioanosa, B. S. *Eos* **72**, 61–62 (1991).
6. Hansen, J., Lacis, A., Ruedy, R. & Sato, M. *Geophys. Res. Lett.* **19**, 215–218 (1992).

Electric traction

LAST week Daedalus invented the converse of a stretch-fabric. Instead of tending to contract in its own plane under an internal surface tension, it tends to expand in its own plane under an internal 'surface compression'. It is woven from conducting wire in such a way that when a current flows, adjacent wires repel each other. The previously slack and flexible fabric then expands tautly, as if it were under tension.

Daedalus intended his new fabric for vacuum-vessels, in which duty its surface expansion could counter the compression of the atmosphere. But such large currents would be needed that only superconducting wire could carry them. So to test the principle, he is weaving more mildly self-tensioning fabrics from normal wire, and is studying their uses.

Already DREADCO tailors have cut out a pair of self-ironing electric trousers. When the current is switched on, the folds and creases of the garment are hauled flat, and it expands taut as if inflated. The way seems open for a military uniform which brings the wearer automatically to attention. Its arms and legs could then be tensioned and relaxed alternately under remote control, thus marching him along semi-automatically at a defined rate. The sergeant-major with the master transmitter could bring parade-ground precision to the most shambling squad of recruits.

In the home, self-tensioning fabrics could be made into self-straightening carpets, self-closing curtains, and even a self-making bed. Hospitals would welcome such a time saver. The matron could remake all the beds of her ward at the flick of a switch, without even removing the patients. Out of doors, rigid self-erecting hoardings should become feasible, and a self-erecting tent needing no guy-ropes or central pole. It will snap skywards when the current is passed through it, and collapse limply when disconnected.

Stressed-skin engineering structures could benefit as well. An electrically pre-stressed skin could carry compressive and shear loads without further reinforcement. DREADCO engineers are planning boats that need no ribs or stringers, tanks and containers rigid without inflation, and aircraft wings and fuselages with no internal bracing. Even the pneumatic tyre could be redesigned as a self-tensioning structure, needing no inflation and unaffected by punctures. But these serious and fiercely stressed applications need such high internal currents that they must await the coming of room-temperature superconductors. David Jones

Moth response to climate

SIR — CO₂-sensitive receptor neurons¹ in the labial palp organ² of the moth *Helicoverpa armigera*, a major agricultural pest, can detect small fluctuations in CO₂ concentration associated with the metabolic activity of food plants with a sensitivity similar to that of modern technical detectors³. As other receptor neurons in insects^{4–6} are strongly temperature sensitive, we would expect that temperature fluctuations, common within the micro-environment of insects⁷, interfere with the detection of CO₂. Instead, we find that the CO₂-receptor neurons in *Helicoverpa* are temperature compensated, albeit only at the CO₂-background levels characteristic of the pre-industrial world.

When a single receptor neuron is exposed to modulation of CO₂ concentration by a constant percentage for a range of backgrounds (Fig. 1a), its response is modulated with a small phase lead³ in synchrony with the stimulus. Both mean action potential rate and response-modulation depth increase monotonically with background. When modulation of temperature is applied instead (Fig. 1b), a different picture emerges: at 100 p.p.m., the response is in phase with the corresponding CO₂ response; at 600 p.p.m. it is

of opposite phase. At 250 and 300 p.p.m., the responses are of the same phases as at 100 and 600 p.p.m., respectively, but are much smaller. This indicates that in this particular neuron exact temperature compensation occurred at a single background, between 250 and 300 p.p.m. The present background (350 p.p.m.) is above the compensation point.

For small signals, the responses are proportional to stimulus contrast³, allowing direct comparison within pairs of measurements at a given background. This enables us to express temperature sensitivity in terms of equivalent CO₂ sensitivity. For example, at 600 p.p.m. (Fig. 1), a CO₂ contrast of 4% causes a response contrast of 59% (gain factor $G_{CO_2} = 14.7$), while 3 K (1% temperature contrast) causes 34% response contrast with opposite phase (gain factor $G_T = -34$). We define the ratio G_T/G_{CO_2} as the temperature sensitivity S , which becomes independent of incidental and often nonlinear properties of individual neurons, such as threshold, adaptation and logarithmic compression.

Figure 2 shows that, with increasing background, S initially decreases steeply from a positive value at low concentrations and becomes negative at higher concentrations, with a decrease in slope. The data closely match the function

$$S = (S_{\infty}c + c_0)/c$$

where c is the background concentration and $S_{\infty} = -4.8$, $c_0 = 1,190$ p.p.m. Therefore, the temperature sensitivity can be modelled as the sum of two terms with opposing signs. Sign and magnitude of the first, background-dependent, term are consistent with the notion that CO₂ must dissolve in an aqueous medium before reaching the molecular sites of sensory transduction within cell membranes. The solubility of CO₂ in water decreases with temperature, by $-2.8\% K^{-1}$ at 300 K. Therefore, if the membrane sites are sensitive to the concentration of CO₂ in the aqueous phase, a value of $S_{\infty} = -8.4$ ($-2.8\%/0.33\%$) is expected, which is within a factor of two of observation. The positive sign of the second, background-independent, term is consistent with observations on other arthropod receptor neurons where temperature sensitivity is determined by the generation of receptor currents⁸. It appears, therefore, that the occurrence of two opposing effects is an inevitable consequence of basic properties of the chemistry of CO₂ and of the physiology of receptor neurons.

We observe that $S = 0$ at a single background, or within a narrow range of backgrounds if we allow for variability between individual receptor neurons. For

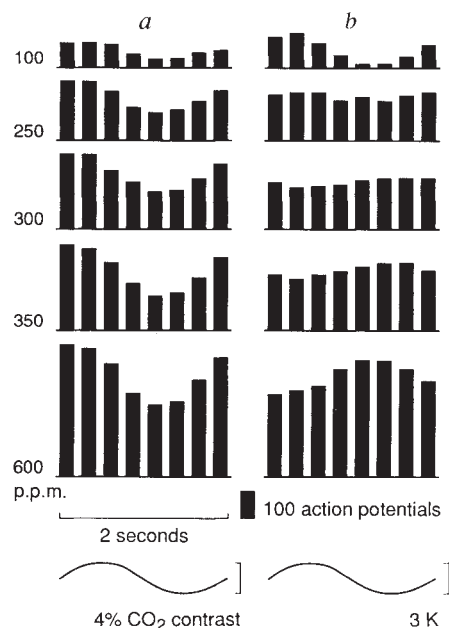


FIG. 1 Pairs of period histograms of action potential probability of a single CO₂-receptor neuron of *Helicoverpa*, logged continuously for 64 periods of a sinusoidal stimulus, in response to modulation of *a*, CO₂ contrast and *b*, temperature, for a range of CO₂ backgrounds (100–600 p.p.m.), at a mean temperature of 298 K. Although the receptor neurons are maximally sensitive at 4 Hz (ref. 3), a stimulus frequency of 0.5 Hz was used because of instrumentation constraints. Measurements conducted at 0.25 Hz gave identical results, indicating independence of stimulus frequency.

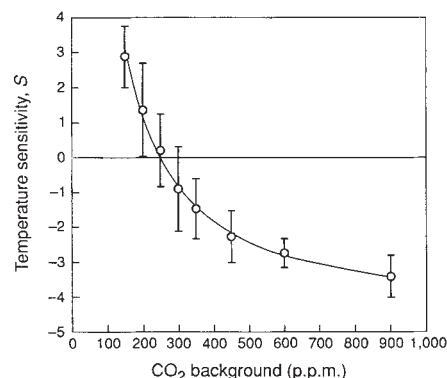


FIG. 2 S as a function of CO₂ background. Circles, means of seven single-cell recordings from different individuals of *Helicoverpa*; error bars, standard deviations; curved line, least-squares fit of the function $S = S_{\infty} + c_0/c$.

the dataset in Fig. 2 and incomplete datasets from *Helicoverpa* and three other lepidopteran species ($n = 19$), all zero crossings occurred within the range 190–320 p.p.m. The present atmospheric CO₂ background is 350 p.p.m.; it was 300 p.p.m. in 1925 and 280 p.p.m. immediately before the industrial revolution. Fossil records from ice-core samples⁹ indicate values of 270–250 p.p.m. between 2,000 and 15,000 years ago; over the 150,000 years before that, the value fluctuated between 190 and 290 p.p.m. Therefore, the moth CO₂-receptor neurons are better adapted to the background levels that prevailed in the recent past than to present levels.

It appears that the unprecedented rate of change in background caused by anthropogenic emissions exceeds the rate at which the moths can genetically adapt. As a result, further increases in CO₂ background will progressively drive the sensory system out of the temperature-compensated range, causing it to confuse fluctuations of temperature with fluctuations of its adequate stimulus, CO₂. There are serious implications of this finding for plant/herbivore interactions in response to future climate change.

Gert Stange*

Chin Wong

Research School of Biological Sciences,
Institute of Advanced Studies,
Australian National University,
GPO Box 475,

Canberra, ACT 2601, Australia

*Also at Max-Planck-Institut für
Verhaltensphysiologie,
82319 Seewiesen, Germany

1. Bogner, F., Boppré, M., Ernst, K. D. & Boeckh, J. *J. comp. Physiol.* **A158**, 741–749 (1986).
2. Kent, K. S., Harrow, I. D., Quartararo, P. & Hildebrand, J. G. *Cell Tissue Res.* **245**, 237–245 (1986).
3. Stange, G. *J. comp. Physiol.* **A171**, 317–324 (1992).
4. Waldow, U. *Z. Verghl. Physiol.* **69**, 249–283 (1970).
5. Coro, F. & Perez, M. *Naturwissenschaften* **77**, 445–447 (1990).
6. Miles, C. I. *Physiol. Ent.* **17**, 169–175 (1992).
7. Willmer, P. G. *Adv. Insect Physiol.* **16**, 1–57 (1982).
8. French, A. S. *J. comp. Physiol.* **A156**, 817–821 (1985).
9. Barnola, J. M. *et al. Nature* **329**, 408–414 (1987).

Recovery of antediluvian DNA

SIR — The recovery of DNA from the tissue remains of extinct animals and its amplification by polymerase chain reaction (PCR) has recently generated much excitement and speculation. Unique information has been gained from mitochondrial DNA sequences of animals extinct for several thousand years, including the giant ground sloth, the moa bird, the sabre-toothed tiger and even the mammoth. Animals which vanished more recently, such as the quagga and the Tasmanian wolf, have also been investigated. The oldest DNA fragments amplified with certainty by PCR are 20,000–40,000 years of age. Beyond this point, time-dependent chemical decay of DNA structure makes successful retrieval difficult¹. A careful investigation² of pyrolysis products of 20-million-year-old magnolia leaves by gas chromatography/mass spectrometry showed that complex macromolecules such as polysaccharides and proteins are completely decomposed, although some lignin remains. It is improbable that any DNA could survive in such material. Despite an initial claim³, a detailed investigation also demonstrated that plant DNA could not be amplified from such leaves⁴.

Two independent reports have appeared on the apparent recovery of DNA, 20–120 million years old, from a single weevil⁵ and from a termite⁶ entombed in amber. This is surprising, because it is unlikely that any DNA sequences could survive that long. Endogenous oxidative damage is one of the major threats to intracellular DNA⁷; 4,000-year-old DNA from dried and mummified tissue has already undergone substantial base oxidation⁸; and amber would not be expected to afford complete protection against oxygen. A major problem with attempts to retrieve very old DNA is the risk of contamination with traces of contemporary DNA. Thus,

DNA might be obtained from pulverised fossil material from the Burgess shale after a sufficient number of PCR cycles, and the DNA might not be identical to any known sequence that has been deposited in data banks, but this does not prove that such DNA was of Cambrian rather than contemporary origin. Furthermore, it is hardly surprising that insect-like DNA can be detected by PCR in experiments carried out in a department of entomology.

The following simple controls should be performed in experiments of this type:

(1) Comparison of positive and negative data. Many attempts to retrieve very ancient DNA by PCR under carefully controlled conditions have been unsuccessful. Such information should be made generally available, for example through a computer bulletin board or the *Ancient DNA Newsletter* (care of Institute of Zoology, Regent's Park, London NW1 4RY, UK), so that negative results can be compared with the occasional positive claims that are published in leading scientific journals.

(2) Reproducibility. Some species of amber-entombed insects have been found many times. It would greatly strengthen the case for credible recovery of very ancient DNA if three such specimens of an insect, collected from different amber deposits and PCR-amplified in different laboratories, were to yield the same DNA sequence.

(3) Negative controls. A piece of amber without an entombed insect, perhaps containing a fragment of wood or stone, should be extracted and PCR-amplified under the same conditions. The DNA amplified by PCR from such material should be compared with that from the insect in amber.

(4) Chemical analysis. The gas chromatography/mass spectrometry technique already used on fossil leaves² should also be used with amber-entombed material. If no proteins or polysaccharides remain, no DNA will have survived either. Similarly, the extremely sensitive ³²P post-labelling technique, originally devised to detect minute amounts of damaged nucleotides in mammalian DNA after exposure to mutagens⁹, could be used to demonstrate the presence or absence of any traces of deoxyribonucleotide material. This method might also yield information on whether most DNA pyrimidines, if detected, would have been oxidised into non-coding ring-fragmented forms (*N*-substituted urea), which could not be copied by a DNA polymerase during PCR.

Recent claims of recovery of 100-million-year-old DNA have overshadowed the valuable and important studies on moderately ancient DNA. Rather than proceed spectacularly further and further back in time with anecdotal

reports on single samples, using the notoriously contamination-sensitive PCR technique, I suggest that the next goal be a convincing report on the amplification of small DNA fragments, say, 100,000 years old.

Tomas Lindahl

Imperial Cancer Research Fund,
Clare Hall Laboratories,
Blanche Lane,
South Mimms,
Hertfordshire EN6 3LD, UK

POINAR REPLIES — Lindahl mentions only two reports on the recovery of DNA from amber inclusions, one from a Tertiary termite⁶ and the other from a Cretaceous weevil⁵. A thorough literature search would have revealed that the first successful extraction was that of a bee in Dominican amber¹⁰ and the most recent is that of a leaf in Dominican amber¹¹.

His 'off the cuff' remark that "it is hardly surprising that insect-like DNA can be detected by PCR in experiments carried out in a department of entomology" shows only that his analysis of the literature is incomplete. In our experiments, none of the extraction, amplification or sequencing was conducted in a department of entomology, or of botany for that matter.

Furthermore, we investigated comparative extant forms only after completely sequencing the extinct species, and we ran more than 20 environmental controls with each PCR cycle. Moreover, experiments were conducted with the extant and extinct species in different places in the building. The controls suggested by Lindahl are quite acceptable and some (reproducibility and negative controls) have already been done for the bee in Dominican amber. Our group has isolated DNA from five separate specimens of *Proplebeia dominicana* and *Rus Hoelzel*, in independent experiments, has isolated DNA from the same bee species in Dominican amber (see ref. 12).

There are many types of fossilization processes, and to assume that the breakdown of DNA is similar in all of them, or is equivalent to that in non-fossilized material, is not scientific. In contrast to Lindahl's statement of the theoretical lifespan of DNA based on experimental data¹, amber maintains the integrity of this molecule for incredibly long periods. We still do not know what components in the resin are responsible for this remarkable degree of preservation, but the fact that DNA is still extractable after 120 million years opens the door much wider for a variety of evolutionary studies. The significance is far more than simply anecdotal.

George O. Poinar Jr

Department of Entomological Science,
University of California,
Berkeley, California 94720, USA

1. Lindahl, T. *Nature* **362**, 709–715 (1993).
2. Logan, G. A., Boon, J. J. & Eglington, B. *Proc. natn. Acad. Sci. U.S.A.* **90**, 2246–2250 (1993).
3. Golenberg, E. M. et al. *Nature* **344**, 656–658 (1990).
4. Sidow, A., Wilson, A. C. & Pääbo, S. *Phil. Trans. R. Soc. Lond.* **B333**, 429–433 (1991).
5. Cano, R. J., Poinar, H. N., Pieniazek, N. J., Acra, A. & Poinar, G. O. *Jr Nature* **363**, 536–538 (1993).
6. De Salle, R., Gatesy, J., Wheeler, W. & Grimaldi, D. *Science* **257**, 1933–1936 (1992).
7. Richter, C., Park, J.-W. & Ames, B. N. *Proc. natn. Acad. Sci. U.S.A.* **85**, 6465–6467 (1988).
8. Pääbo, S., Irwin, D. M. & Wilson, A. C. *J. biol. Chem.* **265**, 4718–4721 (1990).
9. Randerath, K., Leibr, J. G., Gladek, A. & Randerath, E. *Mut. Res.* **219**, 121–133 (1989).
10. Cano, R. J., Poinar, H. N., Roubik, D. & Poinar, G. O. *Jr Med. Sci. Res.* **20**, 619–622 (1992).
11. Poinar, H. N., Cano, R. J. & Poinar, G. O. *Jr Nature* **363**, 677 (1993).
12. Henwood, A. A. thesis, Univ. Cambridge (1992).

First fossil camels from Europe

SIR — Old World fossil camels have long been known to occur in Plio-Pleistocene deposits of Asia¹ and Africa², but have only recently been recorded from Europe. The discovery of camels in Upper Miocene strata at Venta del Moro and Librilla, Spain³ (see figure), raises several questions and provides constraints to hypotheses concerning the phylogenetic roots of Old World camels.

Because of their age and geographical location, the Spanish fossils make previous hypotheses about Old World camel origins obsolete or subject to revision. The possible rates of spread of large mammals (such as camels) colonizing new continents need to be reassessed, and a general revision of Old World camelid fossils should verify previous determinations and fill gaps in the fossil record. In addition, a comprehensive comparison of Old and New World Upper Miocene and Plio-Pleistocene Camelidae may lead to a better understanding of camel phylogeny.

At first glance, the Spanish camels (zone MN 13, Uppermost Miocene, about 7.0–7.5 Myr old) appear to be the oldest known in the Old World, yet geographically they are the furthest from the continent of origin of the Camelidae (North America). This apparent anomaly is undoubtedly partly due to inconsistencies between biostratigraphical scales used in Asia, Africa and Europe. Several of the so-called Lower Pliocene sites of China

and North Africa are in fact Uppermost Miocene, and thus are not greatly different in age from the Spanish sites. But even if the biostratigraphical scales of China and North Africa are revised to account for this, it is likely that the spread of camels to Spain, once they had crossed from the New to the Old World, was rapid, possibly a small fraction of the time represented by zone MN 13.

The absence of fossil camels in collections from other parts of Europe, despite more than a century of research, suggests that camels reached Spain from Asia via North Africa. If this is correct, then camels would have had to cross three 'choke points' in rapid succession in their spread from North America to Spain: the Bering Straits, the Red Sea and the Straits of Gibraltar.

The period of colonization of the Old World by camels was a period of generalized faunal change, with numerous lineages of mammals moving between Africa, Europe, Asia and America. It was also a period of world-wide tectogenic activity (the Rhodanian tectogenic phase)⁴. The lithospheric and biospheric changes during MN 13 are thus likely to be different aspects of the same global process.

Comparison of Old World *Paracamelus*, to which the Spanish fossils belong, with North American *Procamelus*, *Titanotylopus*, *Gigantocamelus* and *Megacamelus* (each of which has been proposed

as the ancestral group⁵) may lead to the determination of the lineage from which Old World camels originated. Age constraints alone suggest that *Titanotylopus* and *Gigantocamelus* of the Blancan to Irvingtonian (4.7–0.7 Myr ago) of North America are unlikely to be ancestral to the Old World camel, unless there were two colonizations of the Old World by camels, an earlier one by *Paracamelus* in the Upper Miocene which became extinct, and a later one by *Camelus* during the Plio-Pleistocene, which still survives. It is more likely that either *Procamelus* (which survived in North America until the beginning of the Hemphillian (about 7 Myr ago) or *Megacamelus* (Late Hemphillian, 6–5 Myr ago) represents the root stock from which Old World camels originated.

Martin Pickford

Geological Survey of Namibia,
PO Box 2168,
Windhoek 9000, Namibia

Jorge Morales

Dolores Soria

Departamento de Paleobiología,
Museo Nacional de Ciencias Naturales,
José Gutiérrez Abascal 2,
28006 Madrid, Spain

1. Schlosser, M. *Abh. Bayer. Akad.* **22**, 1–221 (1903).

2. Stromer, E. *Z. dt. geol. Ges.* **54**, 108–115 (1902).

3. Morales, J., Soria, D. & Aguirre, E. *Estudios geol.* **36**, 139–142 (1980).

4. Scherba, I. G. in *Global Correlation of Tectonic Movements* 49–82 (Wiley, Chichester, 1987).

5. Harrison, J. A. *Smithson Contr. Paleobiol.* **57**, 1–29 (1985).

Powdered moonshine

SIR — Daedalus (*Nature* **363**, 404; 1993) suggested a brilliant (at first sight) idea: to increase the Moon's brightness by dispersing white magnesium oxide powder over its visible surface. This powdering would be achieved by burning magnesium in a rocket whose exhaust plume is directed at the lunar surface.

The economic and ecological advantages could be discussed. First, the averaged moonlight-to-sunlight flux ratio is equal to $A(R/2d)^2$, where A , R and d are the Moon albedo, radius and distance from the Earth, respectively. By increasing the albedo, this ratio can be increased from its present value of about 10^{-7} to at most 2×10^{-6} . This relatively small change in light flux would probably not change the Earth's climate, but the nocturnal metabolism of plants could in principle be affected. Another possible cause for concern is the contamination of near-Earth space by dust particles, as some of the powder would acquire electric charge in the exhaust plume and thus deviate from purely ballistic trajectories.

But in fact there is no need for these concerns. Daedalus's suggested method of powdering the Moon is impossible. The magnesium dust particles will be ejected

at a velocity of a few km per second. Owing to the Moon's gravity, their impact velocity (v) will always be greater than 2.5 km s^{-1} . Because this value is comparable with the velocity of sound in lunar surface material, each dust particle will form a crater. The ratio of the crater volume to the impactor volume can be expressed roughly as $\rho v^2/S$, where ρ is the density and S the maximum shear strength of the lunar surface. Because this ratio is at least 10^3 , nearly all the magnesium powder will be buried under lunar material.

Alexei V. Byalko

Priroda, Moscow, Russia

DAEDALUS REPLIES — Byalko has analysed my scheme most perceptively. However, the magnesium oxide may not all be buried by cratering as he suggests. Magnesium oxide smoke can form particles less than a micrometre across; the grains of lunar regolith are also very small. They will have a shear strength and sound velocity much higher than that of bulk powder. My hope is that the impacting particles will be slowed by a series of individual collisions which may fragment them further, but will not generate a crater that buries them. □



Paracamelus left maxillary molar row (occlusal and labial aspects) and first phalanx (dorsal and volar aspects) from Venta del Moro (zone MN 13), Spain. Scale bars, 5 cm.

Child leukaemia after Chernobyl

SIR — An increase in the incidence of leukaemia has been suggested as an early hazard of the 1986 Chernobyl nuclear power station disaster^{1,2}. This resulted in the exposure to ionizing radiation of 2.3 million inhabitants of the Republic of Belarus, 400,000 of them being children aged 0–14 years. It has been estimated that many received doses in the range 50–60 millisieverts (ref. 3).

The Republican Registry of Blood Diseases in Belarus covers the population of

INCIDENCE OF ACUTE LEUKAEMIA IN BYELORUSSIAN CHILDREN

Radionuclide contamination	Region/City	1979–85	1986–91
Severe	Gomel	35	40
	Mogilev	48	41
Intermediate	Brest	41	40
	Minsk	35	40
Least	Grodno	36	38
	Vitebsk	39	42
	Minsk City	51	48
	All areas	40.7	41.3

Age of children: 0–14; numbers are per million.

the republic, and receives details of cases of childhood leukaemia from haematological and oncological clinics and departments, the Registrar General and autopsies. Diagnoses are based on clinical, cytological (peripheral blood and bone marrow), cytochemical, immunophenotypical and histological data. The registry conforms to international standards of cancer registration. The opportunity has been taken to compare the incidence of childhood leukaemia in Belarus in the period since the Chernobyl accident (1986–91) with that in the preceding period (1979–85).

No appreciable change in the incidence of acute childhood leukaemia in Belarus is seen between the pre-Chernobyl period (40.7 per million) and the subsequent period (41.3) (see table). The same conclusion applies when the sexes are examined separately, males being 47 and 49, respectively, and females 37 and 37, respectively. These rates are not dissimilar to those recorded in many European countries⁴.

There was considerable variation in the level of radionuclide contamination in different regions of Belarus. However, even those regions with the highest levels of contamination, Gomel and Mogilev, showed no increase in the incidence of acute childhood leukaemia in the later period. Indeed this was slightly lower than in the least contaminated areas of the republic (see table).

Our ongoing work, which is part of the

framework of the WHO International Programme on the Health Effects of Chernobyl Accident (IPHECA) and of the European Childhood Leukemia–Lymphoma Incidence Study programme (ECLIS), shows that certain predictions of the leukaemogenic effect of the Chernobyl accident have not yet been borne out. However, as carcinogenesis may involve multiple steps and latency periods may be prolonged, depending on the cancer site, further monitoring of the children exposed to radiation is required.

E. P. Ivanov

G. Tolochko

V. S. Lazarev

L. Shuvaeva

*Research Institute for Haematology and Blood Transfusion,
223059 Minsk, Belarus*

1. Ivanov, E. P. *et al.* *Zdravookhr. Belorussii* **N6**, 57–60 (1990). (In Russian.)
2. Ivanov, E. P. *Atoms Jap.* **37**, 24 (1993).
3. Korolev, V. I. *Zdravookhr. Belorussii* **N6**, 4–7 (1992). (In Russian.)
4. Parkin, D. M. *et al.* *Eur. J. Cancer* **29A**, 85–95 (1993).

Sexual orientation

SIR — In his article¹ on our paper² reporting a linkage between Xq28 DNA markers and male sexual orientation, Maddox conjectures that the Xq28 locus might influence sexual orientation by making the mothers of gay men “over-loving”, in which case “the gene concerned would be strictly irrelevant to the causation of male homosexuality, whose determinants would remain those of nurture rather than nature.” In other words, the gene is postulated to act in women rather than in men.

This speculation violates Mendel’s law of independent assortment. Suppose that there were an Xq28-linked gene X^* that caused women of genotype $X1^*/X2$ and/or of genotype $X1^*/X2^*$ to be “over-loving” and therefore more likely than average to have gay sons. In both cases, all of the sons, either homosexual or heterosexual, would have exactly the same 50 per cent chance of inheriting either the $X1$ or the $X2$ markers. Therefore, in contrast to our experimental findings, no linkage between sexual orientation in males and Xq28 DNA markers would have been observed. One might observe linkage between sisters who have gay sons, but this was not analysed in our study.

Second, Maddox suggests that the study would have been more convincing if we had measured the “incidence of the inheritance of the X-linked markers in the general population”. But such population-based studies require association; that is, a correlation between the trait and a particular allele among unrelated individuals. What we measured

(as should be evident from the title of the paper) was linkage; that is, a correlation between the trait and a locus among related individuals. The differences between association and linkage, and between alleles and loci, are well-established concepts in genetics.

Dean Hamer

*Laboratory of Biochemistry,
National Cancer Institute,
National Institutes of Health,
Bethesda,
Maryland 20892, USA*

SIR — In commenting on the recent report of genetic linkage between markers on the X chromosome and male homosexuality², Maddox proposes the possibility that the putative X-linked gene could play a role in determining whether a mother is “over-loving” rather than contributing directly to homosexual behaviour¹. While this hypothesis is consistent with the observed maternal inheritance pattern of homosexuality, it is clearly contradicted by the linkage data.

If the gene were to affect the behaviour of the mother rather than that of the sons, then brothers, maternal uncles and maternal cousins of homosexuals would still show higher rates of homosexuality than controls as they would be more likely to have had “over-loving” mothers. But because the gene would have no direct effect on the sons, a pair of brothers would be no more likely to inherit the same allele than different alleles from their mother, and hence would show no linkage of the kind observed. It is only when the presence of one allele directly affects male sexual orientation that two brothers would be more likely than chance to share it. If the linkage reported by Hamer *et al.* is confirmed, the gene or genes involved must, regardless of their actual function, influence the behaviour of the males rather than their mothers.

L. Kruglyak

*Whitehead Institute for Biomedical Research,
9 Cambridge Center, Cambridge,
Massachusetts 02142, USA*

1. Maddox, J. *Nature* **364**, 281 (1993).
2. Hamer, D. H., Hu, S., Magnuson, V. L., Hu, N. & Pattatucci, A. M. L. *Science* **261**, 321–327 (1993).

MADDOX WRITES — My original purpose was to draw attention to the interest of a study seeming to be the first identification of a behavioural gene in human beings. I believe it unwise, when breaking such new ground, to settle for the simplest interpretation of the linkage data when two previous such interpretations (published in this journal) have proved unfounded. And should not even molecular geneticists, when announcing new linkages, explain where other research (as on the behaviour of gay men’s mothers) fits in? □

Logic and luck

N. W. Pirie

In Focus, Out of Step: A Biography of Frederick William Twort FRS, 1877–1950. By Antony Twort. Alan Sutton: 1993. Pp. 340. £25, \$50.

ALTHOUGH several distinguished and usually critical scientists were well aware that Twort was an exceptionally imaginative and able research worker, he never gained widespread recognition. One reason was the peculiar structure of the Brown Animal and Sanatory Institution of which he was superintendent for 35 years. It originated from a bequest to the University of London to give free treatment to injured or sick animals, but it had laboratory facilities that had attracted physiologists such as Burdon-Sanderson, Sherrington and Edward Mellanby, and great structural elasticity: that was what attracted Twort. But it had very little money, and it lacked the prestige of a university department. That point was stressed repeatedly in letters from such worldly-wise people as the secretary of the Medical Research Council. A constant search for funding was one of its features.

Twort's personality was another impediment to widespread recognition. Frequent reproof by his tyrannical and ill-tempered father made Twort unduly suspicious — especially of bureaucrats and members of committees that finance, or try to plan, research. Instead of attributing their errors to ignorance and stupidity, he suspected conspiracies and malice, and attacked them. In this biography, his son makes it abundantly clear that, unlike his grandfather, his father was gentle and considerate to his family. He "seldom raised his voice and never lost his temper". His belief "that his children could do anything if they really wanted to, was little more than a reflection of his own enthusiasm".

His early research was on the unreliability of the technique of labelling bacteria by "putting them through the sugars". He regarded many 'species' as varieties or hybrids, and attached great importance to the ability of some cultures, if given sufficient time, to start using substances that they had initially been unable to use. He thought it probable that some pathogens that would not grow in ordinary media would grow if extracts from their 'wild' relatives were added. Attempts to grow leprosy bacilli in media containing extracts from tubercle or Timothy grass bacteria were equivocal, but there was no

doubt about the growth of Johne's bacillus. That causes a wasting disease in ruminants but had, until then, resisted all attempts at cultivation. Diagnostic agents and protective vaccines could not, therefore, be made. Twort's attempts to get finance for developing this technique failed. Possible reasons for the support given to work in other laboratories, along lines that depended on Twort's research,

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Animal house — the early days of the Brown Institution, from *The Graphic*, April 1875.

are discussed in 14 pages of the biography. Obviously, this episode fuelled Twort's suspicions about conspiracies. The need for vitamins, by animals and other organisms, was generally accepted a few years later, and the Johne factor was found to be one of the K vitamins.

Success with Johne's bacillus made Twort assume that viruses did not grow in the media that had been tried, unless active host cells were present, because these cells were needed to supply an essential nutrient. He disregarded the possibility that Beijerinck (1898), Landsteiner (1913), Sanfelice (1914) and others were right in suggesting that viruses needed help from part of the organization of these cells. Thirty years of biochemical

research were needed for general agreement that viruses multiply by deluding the host mechanism into copying them.

Twort assumed that pathogens would usually be accompanied by nonpathogenic variants, that these would have less exacting nutritional requirements than pathogens, might gain pathogenicity during cultivation, and show it when repeatedly applied to a susceptible host. He therefore incubated material containing vaccinia or distemper viruses in a vast range of media, laced with many types of plant extract, and with clay, dung, ditch water and so on. He got no evidence that Berkefeld filtrates contained any more virus than had been put in initially. Occasionally, however, colonies of micrococci grew in filtrates from mixtures that had contained vaccinia. Glassy areas appeared on some colonies. Samples from these areas transmitted the same degeneration to initially unaffected colonies and the agent multiplied in them. This was the discovery of bacteriophage.

Whereas the steps Twort took to discover the Johne factor, and so become one of the pioneers of vitamin research, were logical consequences of his basic assumptions, the discovery of bacteriophage was pure luck. Many bacteriologists must have seen colonies being eroded: Twort also saw that there was something here worth studying. Similar phenomena were soon observed elsewhere and with different microorganisms. Research on phage became, and has remained, a world-wide activity. So has argument about priority in the discovery. About 40 pages are devoted to that theme in this book.

Towards the end of his paper on phage, Twort wrote: "probably it will be found also in cases of dysentery and allied conditions; and I greatly regret that I have not been afforded an opportunity of investigating the dysenteric conditions in the Dardanelles to determine this and other points." His wish was soon met; the army sent him to Salonika in January 1916. There is still no effective antidyenteric vaccine; he was refused permission to try to make an antiserum; and there is no evidence that he tried to control dysentery with a phage. Frustrated by inactivity, and by having his bacteriological competence questioned by military superiors who thought they were dealing with amoebic dysentery whereas he knew it was bacillary, he left Salonika in November 1916. More than 30 pages of the biography refer to this bungle and to the repercussions of his letter in the *Daily Telegraph* on "Science in relation to medicine" which took

Illustrated London News Picture Library

his experience in Salonika as an example.

On returning to the Brown Institution after the war, he made a series of incubators in which virus cultures rotated while intermittently illuminated through prisms and off mineral crystals. In a letter to an editor who was puzzled, as the Medical Research Council was and as the reader will be, about the function of this gallimaufry, he said he was not trying to create life. The design was, however, logically based on his ideas of conditions on Earth when life started. In an article in 1949, and a letter in 1950, he said he had been getting results. In various places he used such phrases as "precellular life", "self-proliferating enzyme" and "living protoplasm that forms no definite individuals". Unfortunately, he did not describe the evidence. It would be a useful guide in planning research along similar lines: Twort's ideas may deserve more attention than they have been given.

Towards the end of his life he wrote hundreds of letters to newspapers and

journals. Most of them were not published. Because of his tendency to mistake foolishness for malignity, they were often strident. Some editors' rejection letters, especially one from the *Lancet*, gave gentle advice: he sometimes took it and achieved publication.

Antony Twort designed the biography as a story about his father, rather than a thesis on his research. Most of it is therefore concerned with family matters and hobbies. So it would be unfair to criticize the results because the scientific material is sometimes unclear: a more systematic system of dating and identifying the many letters quoted would be welcome. There is no complete list of F. W. Twort's publications. In this respect, Twort has been ill-served: the Royal Society obituary does not give a complete list either. □

N. W. Pirie is at Rothamsted Experimental Station, AFRC Institute of Arable Crops Research, Harpenden, Hertfordshire AL5 2JQ, UK.

Through the clouds of obscurity

Robert Hutchison

Stardust to Planets: A Geological Tour of the Solar System. By Harry Y. McSween. *St Martin's*: 1993. Pp. 241. \$22.95.

MOST books on the Solar System cover the Sun, terrestrial planets, outer planets, asteroids and comets in some pre-ordained order. By contrast, "the origin and geologic evolution of the solar system around us" are refreshingly covered in this disordered but unique series of 15 essays. The book will provide new students or amateurs with a broad overview of this complex subject. And it is a joy to read, partly because of the simple language and partly because each chapter is self-contained.

McSween makes it clear that he has not tried to be comprehensive in his overview: he does not wish to swamp the mind of the nonspecialist. Instead, he stops at different objects in the Solar System so as to "focus on some geologic materials, phenomena or processes", thereby killing groups of several different birds with single stones. Thus the second chapter covers the history of astronomy, superposition of strata, photogeology and absolute age determination, all as applied to the Moon. "Hardened hearts" covers not only the "cores and mantles of the terrestrial planets" but seismology, magnetism (or the lack of it), convection (as "in your morning cup of coffee") and the identity of the light element in the Earth's core. In the essay on Mars we are shown that a planet's volcanic activity depends on its size, and learn about atmospheric gases

and the origin of excess radiogenic argon. Fundamental problems and principles are simply and lucidly discussed.

For about 30 years we have known that meteorites contain matter that had 'recently' (4.6 billion years ago!) been processed in stars, but the separation of 'stardust' in the laboratory was achieved only in the mid-1980s. This exciting story is told by the author from first-hand experience: he is actively engaged in the study of chondrites, the meteorites that brought the dust from stars, unscathed, to Earth. McSween uses this chapter to portray the variety of stars and their different fates. Other aspects of astrophysics are dealt with elsewhere: orbital dynamics, the origin of comets, major impacts and the cosmic abundances of the elements are just a few of the subjects discussed using snapshots of different objects in the Solar System. For the origin of life, however, only the Earth will do.

As a North American, McSween tends to use analogues and anecdotes that may be lost on some readers. But the clarity of the scientific prose more than compensates for the US idiom. The figures and diagrams are adequate, but not spectacular (there is no colour). McSween has given us an up-to-date view through the clouds of Venus and the scientific jargon that equally obscure Solar System studies from the nonscientist. Anyone with scientific curiosity will enjoy the tour. □

Robert Hutchison is in the Department of Mineralogy, Natural History Museum, Cromwell Road, London SW7 5BD, UK.

On the edge

Gabriel A. Dover

The Origins of Order: Self-Organization and Selection in Evolution. By Stuart A. Kauffman. *Oxford University Press*: 1993. Pp. 709. \$75, £55 (hbk); \$29.95, £17.95 (pbk).

*Ground Control to Major Tom
Your circuit's dead, there's something wrong
Can you hear me Major Tom*

*This is Major Tom to Ground Control
I'm stepping through the door
And I'm floating in a most peculiar way
And the stars look very different today.*

OUT on the edge of space, David Bowie's Major Tom* hauntingly captured the mood of the 1970s as he cut loose and drifted off. About the same time, Stuart Kauffman took his first steps into the statistical otherworldliness of random parallel-processing Boolean networks as an ensemble approach to locating the 'deep laws' governing biological evolution and ontogenetic unfolding. Twenty years on, Kauffman beckons us, in this massive exposition, to join him on the "edge of chaos" where adaptive complex systems are poised between "frozen" order and "unfrozen" chaos — the supposed global kissing-gate between natural selection and the emergent order of cybernetic systems.

The style is heady, sententious and dangerously seductive, with an audible breathlessness throughout, brought on by the structural wonders visible in nonlinear dynamical systems. It is all too easy to harken to Kauffman's call, to cut loose from our cherished preconceptions of evolution as a hotchpotch of frozen accidents, opportunistically and singularly fashioned by "chance caught on the wing" (to use Monod's phrase). Kauffman is fighting for the conceptual high-ground of inevitable "beautiful and elegant" deep structures, initiated during life's beginnings and "casting an image of permanence and underlying law over biology", the unavoidable "failure" not "success" of selection. Were life to "tap-dance again its erratic course", Kauffman suspects with "quiet passion that below the particular teeming molecular traffic in each cell, lie the fundamental principles of order any life would reexpress".

*Major Tom to Ground Control
Though I've passed 100,000 miles
I'm feeling very still
And I think my spaceship knows which way to go.*

The Central Problem as laid out by Kauffman is that selection acting alone would have an imponderably difficult task in generating ordered, "enormously complex biological systems" simply by shifting around variant alleles in its relentless

*Lyrics by permission of Onward Music Ltd.

search for "arbitrarily located volumes of parameter space", requiring precise genetic control for their subsequent maintenance. The idea here is that the little hyperspace that is adaptively filled is arbitrary, and that such positions could not be reached by selective systems essentially having to frolic up and down adaptive peaks. The assumption is that the actual tree of life, representing all past and present ordered systems, is not simply the result of a historical process, bolting on new bits of genetic information as it opportunistically meandered through the Tree of all Trees, but is primarily a reflection of an ahistorical appearance of self-organizing spots in parameter space spontaneously emerging from huge numbers of interacting genes. On this premise — not far removed from Hoyle's Howler that no self-respecting protein sequence could be randomly assembled from a floating cloud of amino acids by selection — Kauffman confines selection to ensuring maximum "adaptability" on the axes, first, between order and chaos (in the Boolean networks), and second, between "supracritical" and "subcritical" autocatalytic systems, as they arose spontaneously and frequently at life's origins. The central issue is at what point do Kauffman's statistical structures bear upon evolving, historically processed genomes and ontogenies as we know and love them? There are times when the bracing walk through hyperspace seems unfazed by the nagging demands of reality.

Is there truly a "beautiful order which graces the living world" requiring recourse to self-organized properties of large numbers of interacting units? And are we so "overwhelmed with the sheer imponderable complexity of life" that the statistical 'orders' of ensemble theory are our only salvation? In defence of the concept of life as the cookie crumbled, I would answer 'no' to both questions. Developmental and evolutionary geneticists have not yet reached that point of despair at which life has to be understood theoretically and generically from the top down rather than from the bottom up. The 'teeming molecular traffic', directly controlled by very local and combinatorially acting gene regulatory networks, can be attacked reductionally and diachronically without recourse to the generic "laws of form" poking out of cybernetic networks. The evolved modular structures of genes and organisms point to manageable simplicity rather than hopeless complexity soliciting an ensemble approach.

The first foray into ensemble theory concerns the NK model of adaptive landscapes. Before Kauffman, models of adaptive landscapes were confined largely to N number of genes being independent units with additive effects on the fitness of

genotypes. Kauffman investigates the distribution of genotypes and fitnesses by assuming (realistically) that genes interact with each other. The average K number of interactions can vary from zero to $N-1$, but only at $K=2$ (with randomly assigned fitnesses) is a "correlated rugged landscape" achieved, in which variant genes, with short mutational distances from wild type, can explore very high neighbouring adaptive peaks rather than being trapped on low local peaks and suffering the "complexity catastrophe" associated with high K . Kauffman concludes that such $K=2$ landscapes "undoubtedly underlie adaptive evolution at the molecular and morphological levels". $K=2$ landscapes

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Hyperspaceman — Stuart Kauffman.

also explain *en passant* radiation and stasis. Von Baer's laws, the Cambrian explosion and Permian quiescence, homoplasia and convergent evolution, 'evolvability' and sustained fitness. Be that as it may: the primary interest is in the magic number $K=2$, which also emerges as the nexus of all things biological in the Boolean networks.

Boolean networks are logical on-off switching systems. On the assumption that genes are on or off as a consequence of other genes being on or off, Kauffman considers thousands of genes at which K connectivity and Boolean functions (for example, 'Or' and 'And') are thrown in at random. When $K=2$ there is a critical space between 'frozen' order ($K<2$) and 'unfrozen' chaos ($K>2$) at which biological complexity is poised. At this 'edge of chaos', network landscapes are again rugged and correlated, providing selection with an easy route to higher things. The short networks of interactions when $K=2$ are "attractors" which,

according to Kauffman, represent cell types with the total number of alternative attractors being equal to $\sqrt{N}$. So a human genome of about 100,000 genes would give rise to roughly 300 cell types. Biological complexity (as measured by cell-type number) is therefore governed by the richness of the couplings (K) between a large number of genes, hence satisfying "the rational morphologist's dream of the laws of form".

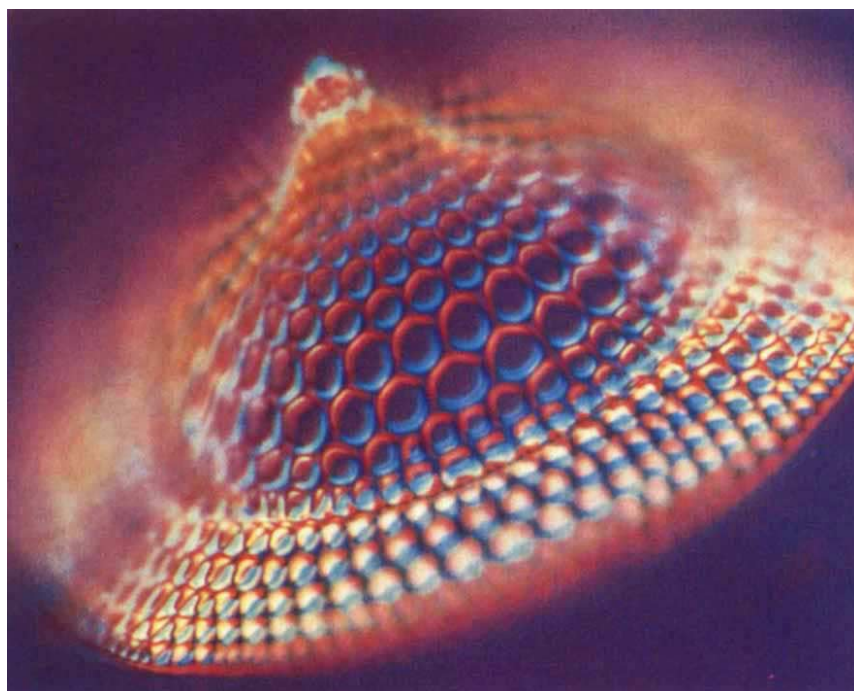
The problem with all this is that although it is logically possible that limited state-cycles emerge as generic properties of complex parallel-processing Boolean networks (in particular when both biases and forcing 'canalizing' structures are thrown in to limit the quality of the K connections), such state-cycle attractors are not meaningful in a functional sense (that is, in the sense understood by the great unwashed masses of biochemists). And here lies the rub.

In order to circumvent this minor irritation, Kauffman elaborates further models that produce more meaningful networks involving catalytic peptides and enzymatic RNAs, as they emerged at the dawn of life. Such autocatalytic closures, representing the first fumbblings towards a coordinated web of metabolism, are spontaneously self-reproducing, not needing a self-replicating 'nude' genome. At first they are 'supracritical' (tolerant, flexible and soliciting any available energy source) but move, via selection, to being 'subcritical' in order to achieve greater specificity and efficiency. This sounds reasonable, if unproven. For Kauffman, the case is "overwhelming, and the argument is insensitive to most of the details". (Unfortunately, God lies in the details.) Hence why, having attained some meaning, life then moved on to meaningless Boolean-generated genetic control circuits is not clear (one step forward, two steps back). Kauffman attempts to provide a link between meaningful autocatalytic closures and meaningless NK state-cycles by use of random grammar models which, he senses, lie at the heart of just about everything. These account for particular couplings among symbol strings and provide "internal meaning" to statistical connectivity whether in biology, economics, culture, psychology or "scientific ideational". A bit of applied string theory to this section might have improved intelligibility, but talk of Jets, Lightning Balls, Mushrooms, Fixed and Traveling Eggs, Filigreed Fogs and Pea Soups imparts little meaning to those not primarily engaged in such exhilarating but seemingly Jolly Lunacy.

*This is Ground Control to Major Tom
Your circuit's dead, there's something wrong.*

Fortunately, and rather surprisingly, in the final 200-page section dealing with cell differentiation and morphogenesis,

University of Pennsylvania



MICROFOSSIL — glassy opal skeleton of a unicellular marine organism, taken from *On the Nature of Things* (Aperture, \$40). This beautiful volume commemorates the work of the influential science photojournalist Fritz Goro (1901–1986), most famous for his 50-year collaboration with *Life* magazine. The book's 159 technically brilliant images are accompanied by comments from eminent scientists who worked with Goro.

Kauffman appears to abandon parallel-processing ensemble models either by marginalizing most of their effects through the interesting device of “forcing structures” (leaving only a small subset of genes actually responsible for pattern regulation) or by introducing gene-independent Turing-style reaction-diffusion models. When he holds up *Drosophila* development as an example of such phenomenology, he is forced to recognize that there is little that is not under the sequential control of small numbers of genes acting in very local patches of cells in bizarre, combinatorial and often redundant operations that can be explained as much by history as by ensemble or gene-free ‘field’ theory.

For example, Kauffman maintains that most genes are regulated by only a few other genes ($K=0\rightarrow6$) via canalizing Boolean functions. If, for example, there are half-a-dozen *trans*-acting controlling factors, any one of which can be missing to turn a gene off, then the system is canalized. The Boolean ‘Or’ function is clearly canalized. Hence most genes (the house-keeping genes?) fall into chains of command structure in which genes are fixed in either active or inactive states. These represent the “frozen cores” by which each cell type supposedly becomes uniquely different. But to provide flexibility, a few genes, by virtue of their low connectivity ($K=1$), fall into “independent feedback loops” with alternative states, which through a “binary combina-

torial epigenetic code” can also give rise to novel cell types. Then (we’re nearly finished), if the products of a subset of genes can diffuse between cells (as they do), the tissue as a whole can settle down to being an “overall dynamic attractor”. By tweaking the parameters in such ways, Kauffman is dangerously close to simply restating in cyberneticspeak what we actually know is going on in, for example, the hierarchy of regulatory decisions, from *bicoid* to the segment-polarity genes that control segmentation during early *Drosophila* embryogenesis. Similarly, having shown that the original Turing models of ‘standing waves’ (the result of local differentials in rates of reaction and diffusion) are insufficient to account for embryogenesis, Kauffman ends his account of the origins of order by proudly wheeling on his and Brian Goodwin’s ‘Four Colour Wheels Model’ by which cells take their cues from phase differences of gene-encoded morphogens having progressively shorter wavelengths.

This is not the place to examine the extent to which the local protein and RNA gradients produced by the maternal, gap, pair-rule and segment-polarity genes can be squeezed into the bifurcating system underlying the Four Colour Wheels. The irony lies more with the point that all such diffusing molecules are *gene-controlled*. Morphogen concentrations and locale take their cues from the intimate and increasingly bizarre local interactions

involving redundant *cis*-acting binding sites, differential concentration-dependent binding affinities and intimate cell-cell interactions, and not obviously from the global dimensions of the tissue, as required by field models. Once we ‘reduce’ morphogens and pattern demarcation lines to the combinatorial genetic processes (the ‘syntagmata’ of Garcia-Bellido) of a finite number of genes whose products diffuse or signal to neighbouring cells, then we end up, after 700 pages, treading the same realistic stomping ground beloved of Earth-bound developmental geneticists. Such finite regulatory networks and local cell autonomy do not necessarily require the spontaneous self-organization of Boolean systems or pattern-generating morphogenetic fields but, rather, with their bolt-on modularity, are more easily understood as the accidental but functional products of an incessant process of evolutionary molecular Lego.

With its imagery of adaptive landscapes, morphogenetic fields and restrictive canalizing functions, this attempt at computerized biology is strangely old-fashioned. Importantly, it does not embrace the evolutionary consequences of genetic and functional redundancy and genomic flux. Such features impart a degree of internal tolerance and the facility for molecular coevolution in combinatorial regulatory systems that is far removed from the hard lock-and-key imagery of fixed adaptive landscapes and canalizing *K* connectivity. Landscapes of evolved organisms are more akin to the heaving surface of the ocean due to genomic and ecological turbulence, being both adaptive and nonadaptive, but functional for all that. Designing life through ensemble theory is as cosy as designing a jumbo jet on a computerized drawingboard. The real trick of *Life* is the gradual evolution of a jumbo from the first wire-and-wood contraptions while the bloody things are still flying in the air! This requires the internal tolerance governing evolving molecular interactions that is provided by modular and redundant systems.

It may be that Kauffman is not so much ahead of his time as that most of us are behind ours. *NK*, as an explanatory model for the origins of order, may come to signify Not Korrekt or Nobel Kauffman. My head, full of awkward biological knowledge, tells me the former might be the outcome, but my heart, incessantly fighting the vulgar Darwinism of solitary selfish genes, would settle for the latter.

*This is Ground Control to Major Tom
You've really made the grade
Now it's time to leave the capsule if you dare.* □

Gabriel A. Dover is in the Department of Genetics, University of Leicester, Leicester LE1 7RH, UK.

Building the crust at the Mid-Atlantic Ridge

Deborah K. Smith & Johnson R. Cann

The morphology of the sea floor at mid-ocean-ridge spreading centres provides a key to understanding how ocean crust is constructed. Images of the axial zone of the slow-spreading Mid-Atlantic Ridge, obtained at a range of spatial scales, show that crustal construction is complex and highly variable, reflecting an underlying three-dimensional variability in magmatic, tectonic and hydrothermal processes.

At mid-ocean ridges magmatic, tectonic and hydrothermal processes combine to form the oceanic crust. The surface expression of this complex system is distinctive sea-floor topography, the characteristics of which vary from ridge to ridge, and are primarily associated with spreading rate¹⁻⁴. At fast-spreading ridges such as the East Pacific Rise (up to 180 mm yr⁻¹ full rate) the topography is distinguished by an axial high 1–2 km wide^{3,5}. Fault scarps are of the order of tens of metres high, and low-relief lava flows dominate the constructional volcanism. By contrast, most slow-spreading ridges such as the northern Mid-Atlantic Ridge (~25 mm yr⁻¹ full rate) are dominated by an axial rift valley 30–45 km wide and 1–2 km deep⁶. This 'median valley' is composed of an inner valley floor 5–12 km wide, bordered by valley walls in which numerous normal faults, some of which are hundreds of metres high, displace the crust upwards to form the crestal mountains. The inner valley floor is the primary site of crustal construction⁷. Eruptions on the valley floor commonly build prominent axial volcanic ridges ranging in size up to several hundreds of metres in relief, 1–5 km wide and several to tens of kilometres long^{8,14}. In most places, axial volcanic ridges are built of piled-up and coalesced individual small (1–2 km in extent) volcanoes^{9,10,13,14}.

The narrower axial zone and simpler structure of the East Pacific Rise (EPR) can be imaged in a single multibeam bathymetry swath. Because of the ease of mapping the EPR, long sections have been surveyed, and the nature of ridge segmentation and offsets between segments, as well as intrasegment variability in ridge crest topography, were first understood there¹⁵. Classic models of the structure of the ocean crust are based on investigations of the EPR, where it is thought that a single magma chamber extends along sections of the axis and a sheeted dike complex lies above it.

Until recently, because of the wide axial zone at the Mid-Atlantic Ridge (MAR), there was no equivalent mapping in the Atlantic. The main focus of investigations at the ridge axis was to understand the two-dimensional evolution of the median valley. In the last few years, however, there have been rapid advances in along-axis bathymetric coverage. As surveys of more spreading segments of the MAR are obtained, the three-dimensional variability in topography and ultimately ridge-crest processes at a slow-spreading ridge is being illuminated.

Here we combine insights gained from early studies of small sections of the ridge axis of the northern MAR with those gained more recently from mapping and imaging numerous spreading segments, to build on existing models for crustal construction at a slow-spreading ridge. The structure of the shallow crust at the MAR is probably fundamentally different from the classic model for the faster spreading EPR. At the MAR no steady-state magma chamber exists beneath the ridge. Instead, discrete magma pockets rise through the shallow crust to feed small axial volcanoes throughout the inner valley floor. The small axial volcanoes may be isolated or form linear chains, and often pile on top of each other and coalesce to build larger axial volcanic ridges. As crust is built and spreading continues, new inner-valley-floor bounding faults propagate into the valley floor often intersecting an axial volcanic ridge. Thus, part or all of the vol-

canic ridge is destined to be incorporated into the valley walls. Based on the observed topographic variability, sections of an individual spreading segment behave independently. This suggests that entire segments do not go through evolutionary cycles, although sections of the ridge may be more magmatically or tectonically active at certain times.

Historical perspective

Our view of the topography at the northern MAR has improved significantly in the last 30 years. The gross structure of the MAR was known in the 1950s and 1960s from bathymetric profiler records^{2,16,17}. These included profiles collected on random tracks across the ridge, as well as a few detailed surveys such as at 45° N (ref. 17). The echograms revealed a wide and deep median valley running along the axis. From magnetic, gravity and bathymetry data it was recognized that new oceanic crust was generated in the median valley.

The first detailed picture of the fine-scale topography of the axial zone of the MAR was provided by Project FAMOUS (French-American mid-ocean undersea study), a multi-disciplinary programme that included more than 25 cruises and dives at the axial zone near 37° N (ref. 18). Nested surveys were designed in which images of the sea floor were collected in smaller patches at increasingly finer scales. Other significant morphological investigations of the axial zone at the northern MAR were undertaken at the AMAR (Alvin Mid-Atlantic Ridge) area¹⁹ the adjacent segment to the south of the FAMOUS segment, the TAG (Trans-Atlantic geotraverse) area²⁰ at 26° N and the MARK (Mid-Atlantic Ridge at Kane) area¹⁰ at 23° N

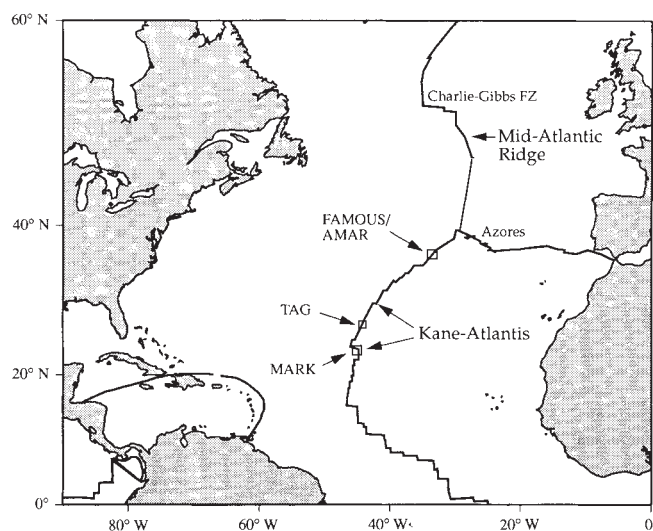
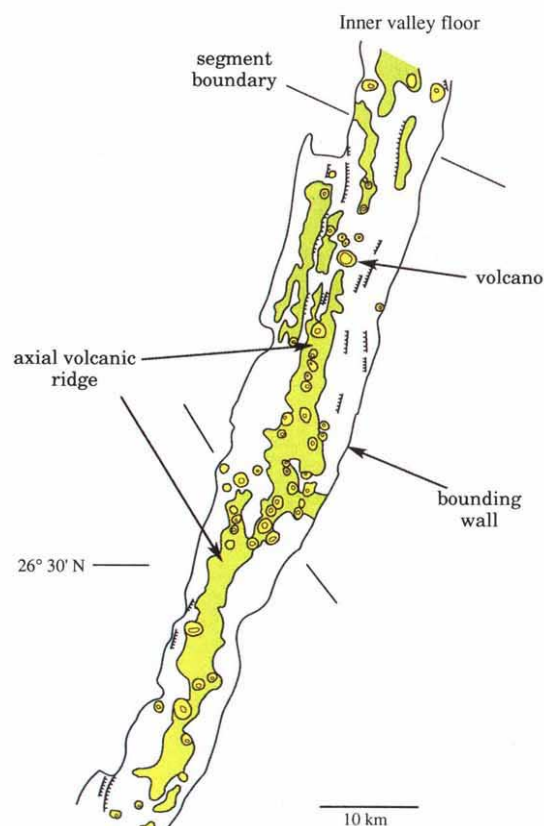
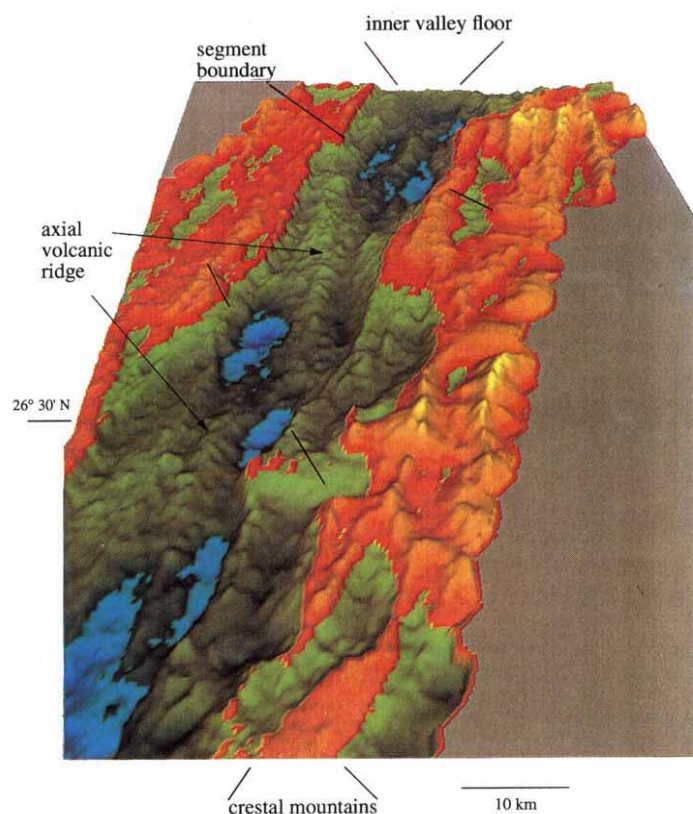
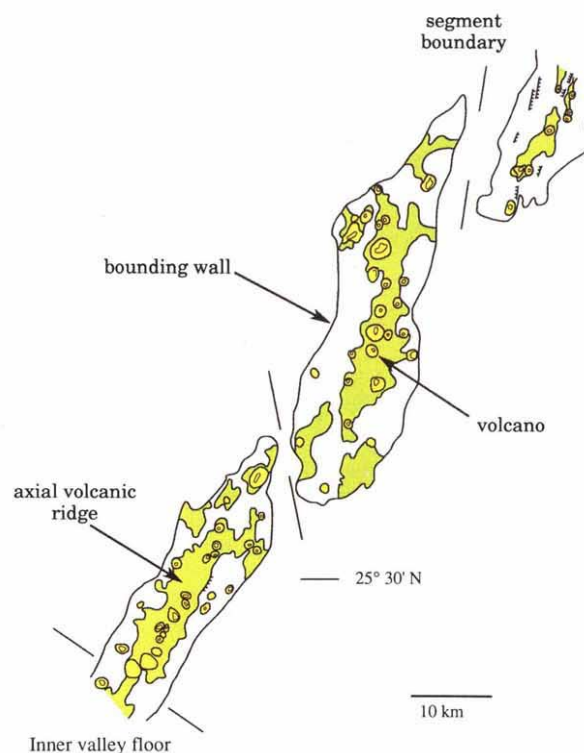
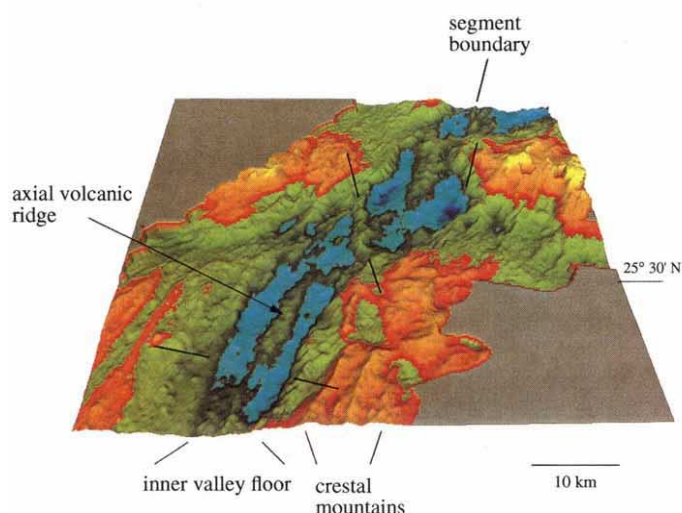


FIG. 1 Location map showing the northern Mid-Atlantic Ridge and the sites of detailed multibeam bathymetry, side-scan sonar, and submersible surveys discussed in this review.

(Fig. 1). These studies were concerned mainly with understanding the variability in ridge-crest processes through time. Models were proposed for crustal accretion in which eruptions on the inner valley floor constructed volcanic ridges such as Mt Venus in the FAMOUS area; the volcanic ridges were then faulted and carried out of the median valley and new ridges were built^{8,9}.

Along-axis mapping of the MAR was limited, however. Multi-beam bathymetry data had been collected for two spreading segments at FAMOUS²¹, and two at MARK¹⁰. Each of these four spreading segments has distinctly different topographic characteristics (for example, size and distribution of volcanic constructions and faults). This made it difficult to determine if



there was a 'typical' spreading-segment morphology, whether the four segments represented different stages in a magmatic/tectonic evolutionary cycle, or whether they represented parts of a more complex process.

A few years ago multibeam swath bathymetry data were collected over approximately 900 km of the ridge between 24° and 31° N (from the Kane transform to just north of the Atlantis transform)²². More than 18 spreading segments were mapped with coverage extending laterally from the inner valley floor to the crestal mountains (Fig. 1). More recently, within the Kane–Atlantis region, the inner valley floors of nine of the spreading segments were imaged using a deep-towed side-scan sonar system (D.K.S. and J.R.C., unpublished data). These surveys reveal the wide variability in along-axis morphology, suggesting that crustal accretion at the MAR is a complex three-dimensional process.

Topography of the inner valley floor

The topographic characteristics of the northern MAR from studies at FAMOUS, AMAR, TAG, MARK, and the Kane–Atlantis areas (Fig. 1), form the main constraints on our model for building the crust. These include the larger-scale (tens of kilometres) structure of individual spreading segments, as well as the smaller-scale (1–2 km) volcanic edifices that are formed throughout the inner valley floor and their relationship to the segment-scale axial volcanic ridges. Inner-valley-floor faulting and fissuring, and initiation sites of valley-floor-bounding scarps provide insight into tectonic processes that modify the newly formed oceanic crust as it moves from the axial zone.

Large scale features. Most of the northern MAR is composed along its length of discrete spreading segments tens of kilometres long^{9,11,12,23–25} (Fig. 2). Segmentation of the ridge is revealed by

the lateral offset of the axis of volcanic activity by transform faults and non-transform discontinuities. Offsets range from 3 to 50 km (ref. 12), separating spreading segments that may have very different volcanic and tectonic styles. Segmentation of the ridge is also seen in the regular change in depth along the axis: depths are shallow near the centre of a segment and increase at the ends¹². In addition, variations in seismic crustal structure along the axis gives evidence for segmentation²⁶. Also, mantle Bouguer gravity anomalies suggest that magmatic accretion is focused at distinct centres that correspond to individual spreading segments identified from the bathymetry^{27,28}. It is inferred

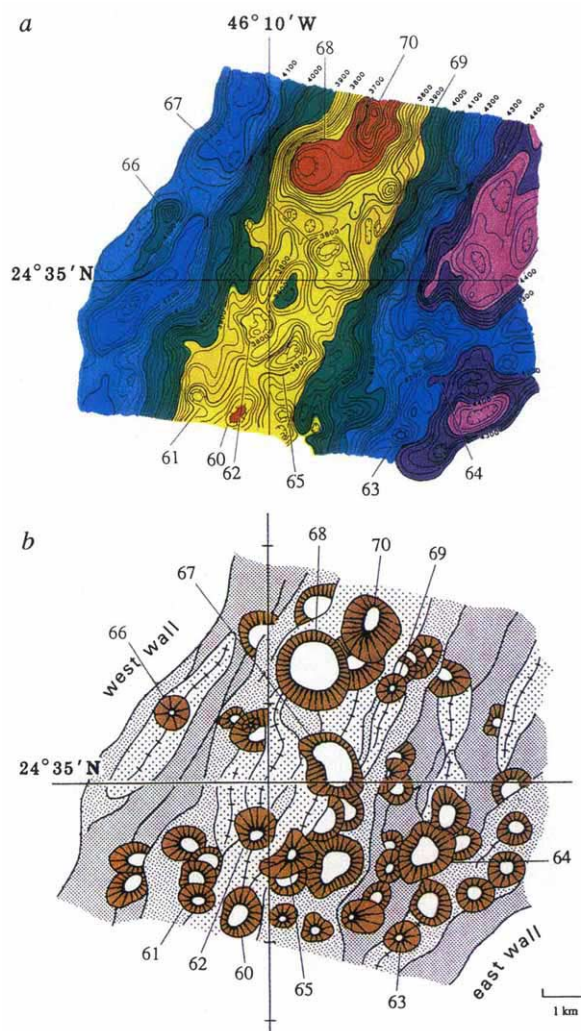


FIG. 3 a, Sea Beam bathymetry²² of a portion of the inner valley floor plotted at 20-m contour intervals. Tick marks point downhill. An axial volcanic ridge and flanking deeps make up the topography of the valley floor. The shallowest portion of the ridge is shown in yellows and reds; purples indicate the deeps. Numbers show the locations of near-circular volcanoes >50 m in relief identified by Smith and Cann^{13,14}. b, Geological interpretation of the Sea Beam bathymetry. Circular volcanoes are shown with summit areas white and slopes hatched and shaded brown. Volcanoes numbered in a are also numbered here. Others depicted are <50 m in relief, the minimum height specified for volcano identification^{13,14}. Contours that are rounded to form sections of circles are interpreted to be volcanoes buried by later volcanism or volcanoes abutting pre-existing features. Ridges and slopes with parallel contours are interpreted as hummocky flows from fissures; these are shown stippled and overlapping one another with lighter stippling denoting the youngest eruptions. Straight lines with ticks mark the crest of the ridges. It is inferred from the Sea Beam bathymetry that small volcanoes and flows coalesce to build the axial volcanic ridges.

FIG. 2 Sea Beam bathymetry²² (left), gridded at 150-m spacing, of the inner valley floor and crestal mountains of the Mid-Atlantic Ridge (image produced at Boise State University by M. Dougherty and S. Garland using AVS and Matlab). Blues represent depths >4,000 m; greens, 3,000–4,000 m; reds, 2,000–3,000 m; and yellows <2,000 m. Multi-beam echosounders typically resolve well features >200–300 m in diameter. Also shown (right) is the geological interpretation of the topography of the inner valley floor from Sea Beam maps contoured at 20-m intervals. On the interpretation the bounding scarp is shown as a continuous line. It is not a single fault, however, but jumps from scarp to scarp along a segment, widening and narrowing the valley floor. Segment boundaries are indicated by a broken line across the ridge. The axial volcanic ridge is shaded green, and circular volcanoes identified from the Sea Beam data^{13,14} are in yellow. Near-circular small volcanoes are found throughout the valley floor, and are typically associated with the axial ridges. Axial ridges are the large-scale volcanic building blocks of the topography on the inner valley floor. Small volcanoes pile upon one another to build the ridges and their flanking deeps. *Top*, The median valley between 25° 15' N and 26° 00' N. Two complete spreading segments are shown. Each segment contains a large axial volcanic ridge which is, in general, centrally located on the wide inner valley floor. The axial ridge becomes discontinuous towards the ends of the segments, and in places is linked to the bounding scarp. *Bottom*, The median valley between 26° 15' N and 27° 10' N. Two axial volcanic ridges overlap at the segment boundary in the centre of the figure; there is a volcanic connection between the two ridges. Note the changing morphology of the axial volcanic ridge; its orientation fluctuates, in places it becomes discontinuous and it is linked to the bounding scarp. The major bounding fault on the west (upper left in the figure) extends towards the south into the inner valley floor as minor faults cutting and splitting an axial ridge. Presumably the inner-floor-bounding scarp will jump inward to the site of these faults, thereby incorporating a large section of the floor and axial ridge into the walls of the median valley. A volcanic ridge more centrally located within the inner valley floor is under construction; the feature labelled 'volcano' is part of this ridge. The primary locus of extrusive eruptions may be shifting laterally to build this ridge.

from the gravity data that the oceanic crust is thicker or mantle temperatures higher (or both) at segment centres than at segment ends.

The edge of the inner valley floor can be defined structurally as the base of the first major scarp of the bounding walls. The edge cuts across bathymetric contours, and steps from scarp to scarp along the length of an individual spreading segment, as scarps are not continuous for the whole segment length^{9,12,29}. Average inner-valley-floor widths vary between segments, ranging from 5 to 12 km (ref. 12). Floor widths also vary along the length of a segment by a few kilometres as individual bounding

scarps step inward.

Most segments contain within the inner valley floor prominent axial volcanic ridges that are the principal sites of lava extrusion^{8,9,11,14,30-32} (Fig. 2). An axial volcanic ridge may be centrally or asymmetrically located between the bounding scarps. Though often linear, the orientation of an axial volcanic ridge may meander. Its trend may differ from the strike of a bounding fault or of the segment as a whole, and some ridges terminate at a bounding scarp. Axial volcanic ridges are commonly discontinuous along their strike, and in addition, two or more subparallel and adjacent axial volcanic ridges can coexist

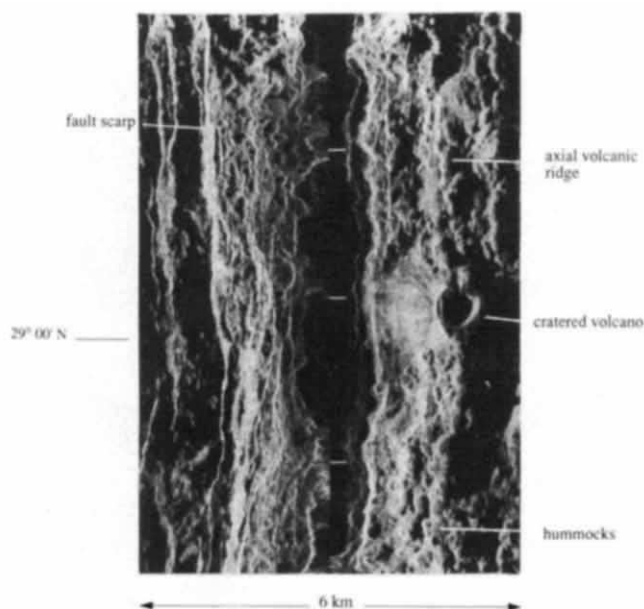


FIG. 4 TOBI (towed ocean bottom instrument) deep-towed side-scan sonar image of the western edge of the inner valley floor of the MAR near 29° N (D.K.S. and J.R.C., unpublished data). The image is slant range corrected, and plotted assuming a flat bottom. The vehicle track is in the centre of the image. Half-hour tick marks are shown along the track. No data are returned from directly below the vehicle. In general, the data in a strip ~1-km wide centred directly below the vehicle are unreliable. The pixel size is ~10 m; features with diameters >50 m are well resolved by the TOBI instrument. The TOBI vehicle is towed 400–600 m above the sea floor and images a swath 6 km wide. The frequency of the side scan is 30–32 kHz. Bright is a reflection, dark is a shadow. The right side of the image shows an axial volcanic ridge. A large (~220 m high) flat-topped, cratered volcano makes up part of the axial ridge. The diameter of the top is ~600 m. Hummocks and hummocky ridges cover the other parts of the ridge shown. (We use 'hummock' to mean a rounded mound 50–500 m in diameter with relief <50 m.) In the upper section of the figure the volcanic ridge is cut by small faults and fissures. The left side of the image shows the valley-floor-bounding scarp. It is composed of several scarps with different relief; the scarps split and merge along the strike of the segment.

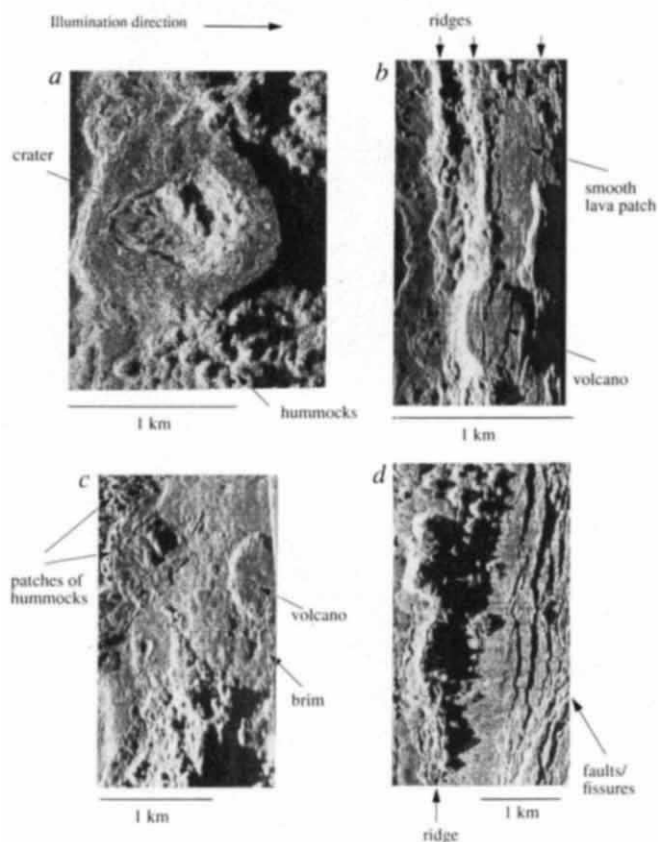
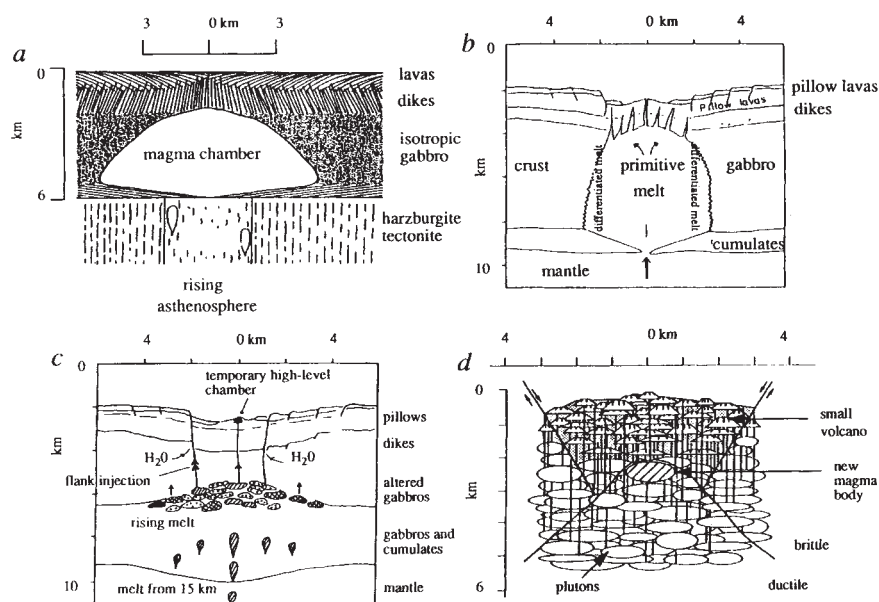


FIG. 5 TOBI side-scan images of small volcanoes that build the topography on the inner valley floor of the northern MAR (D.K.S. and J.R.C., unpublished data). See Fig. 4 for TOBI data characteristics. All features are illuminated from the left. It is not possible to determine flow morphology from these images, so we cannot discriminate between sheet flows and pillows. Therefore, in the following descriptions the term 'smooth' refers to the texture of the feature on the TOBI images. *a*, Smooth-textured, flat-topped volcano with summit height of ~70 m, a basal diameter of ~1,600 m and a summit diameter of ~1,100 m. The summit crater is filled with what appear to be hummocky flows (see Fig. 4 for definition of hummocks). *b*, Flat-topped volcano ~120 m high with basal diameter of ~1,350 m, and summit diameter of ~625 m. A low relief, smooth lava flow which may have erupted from the base of the volcano is adjacent and to the north of the volcano. Also observed are hummocky ridges and patches of hummocks. *c*, Hat-shaped volcano. The central dome of the volcano is at the upper right and is encircled by a smooth textured brim of lava. The basal diameter of the entire feature is ~2,100 m and it has a summit height of ~80 m. The diameter of the central dome is ~650 m. Hummocky patches, and additional smooth patches of lava overlap the brim of lava. *d*, Hummocky ridge ~3,300 m long, 400 m wide and 30 m high. The ridge is surrounded by smooth, low-relief lava flow. The orientation of this small ridge follows that of neighbouring small faults and fissures, and is parallel to the axis of the segment.

FIG. 6 Models for the construction of the oceanic crust. *a*, Melt rises continuously to feed a large high-level magma chamber beneath the ridge. It is erupted through dikes creating a sheeted dike complex leaving behind cumulates to form isotropic gabbros as they freeze (from ref. 72). *b*, Melt rises to accumulate in a magma chamber similar to *a*, however, the magma chamber is zoned laterally producing petrologically different lavas at the edges of the inner valley floor as compared to those at the centre of the floor (after ref. 36). *c*, No large magma chamber exists. Instead small batches of melt rise through the mantle and accumulate at the base of the elastic crust. Each magma chamber erupts separately through cracks in the overlying brittle carapace; melt rises beneath the entire width of the valley floor so that flank eruptions can occur (from ref. 48). *d*, As in *c* no large magma chamber exists. Magma bodies rise to the brittle-ductile transition within the crust. They shoulder aside and depress older plutons, and erupt through dikes and pipes to build volcanoes (hatched) and hummocky flows (stippled) (see Fig. 4 for definition of hummocks). Each small volcano and hummocky flow is assumed to be fed from a separate small magma body. Therefore the lower crust is made up entirely of small plutons (from ref. 14).



within a segment. In some places, eruptions have generated volcanic edifices that trend laterally across the valley floor and link the axial volcanic ridge and the bounding wall.

In the spreading segments mapped at the northern MAR, axial volcanic ridges have a wide spectrum of styles. For example, at FAMOUS, Mt Pluto and Mt Venus (both 100–200 m high, about 3 km long and 0.5–1 km wide) form part of a discontinuous axial volcanic ridge that runs along the valley floor⁸. Examples of axial volcanic ridge morphology from the Kane–Atlantis area are shown in Fig. 2. Figure 2 (top) shows the bathymetry data and geological interpretation for the axial zone of two spreading segments separated by a non-transform offset. Both segments have a robust axial volcanic ridge on the inner valley floor. Each volcanic ridge is approximately 30 km long,

3–6 km wide and 500–600 m in relief; at the ends of the segments the ridges are discontinuous and commonly linked to the edges of the valley floor. In Fig. 2 (bottom), bathymetric data and geological interpretation are shown for two additional spreading segments. The axial volcanic ridge in both of these segments is on average 2–3 km wide and 150–200 m high, although each ridge widens and narrows along its length. Moreover, in places the volcanic ridge is discontinuous, faulted into several sections, and linked to the bounding scarp.

The widespread occurrence of axial ridges on the inner valley floor of the northern MAR indicates that they are the primary large-scale volcanic building blocks of the topography in the axial zone. Volcanic ridges with similar morphology to those observed on the inner valley floor have been imaged in the steps

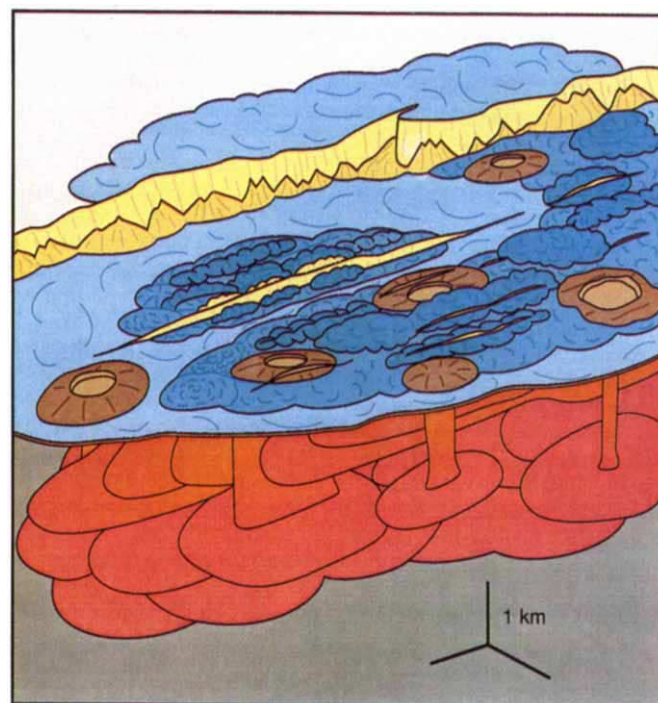


FIG. 7 Schematic cut-away perspective view of part of a spreading segment of the northern Mid-Atlantic Ridge. The cartoon illustrates the first crustal block of the bounding walls, the bounding scarp with debris fans at its base, the topography of the inner valley floor, and the small magma bodies feeding the ridge. Scale bar shows one kilometre; there is no vertical exaggeration. Detailed volcanic ornament shows the now active axial volcanic ridge sections, with links to the bounding wall scarp, and a recently faulted section of an axial volcanic ridge. The inner floor is cut by small faults and fissures. The larger fault cutting the axial ridge is the site of a new bounding fault within the inner valley floor. Sketched volcanics indicate the flanking deeps of the axial ridges on the inner floor. They also indicate volcanic morphology on top of the first faulted crustal block of the median valley walls. The lower part of the figure is a view below the sea floor showing dikes and pipes of different geometries rising from small flat magma bodies. The different volcanic morphologies on the sea floor represent the different building blocks of the topography seen in deep-towed side-scan sonar images and include circular volcanoes and small ridges that pile up to form the axial volcanic ridges.

of the bounding walls^{8,9,30,31,33} and it has been inferred that they are fossil axial ridges. If this is the case, the building of axial volcanic ridges has been the preferred style of volcanism over at least the past million years.

Small-scale volcanic building blocks. Individual volcanic edifices dominate the small-scale (1–2 km) morphology of the inner valley floor of the northern MAR. This is very different from the fast-spreading EPR, where low-relief lava flows dominate the constructional volcanism^{3–5}. For example, multibeam data indicate that small volcanoes are abundant at the MAR between the Kane and Atlantis fracture zone where more than 450 near-circular volcanoes in the height range 50–200 m were identified on the inner valley floor^{13,14} (Fig. 3). Small volcanoes are distributed over the valley floor of all spreading segments surveyed. They may be isolated or form chains with linear trends^{14,30,31}, but in general, are associated with the larger-scale axial volcanic ridges^{12,14,30,31}. In some segments, chains of volcanoes mark the crests of the axial volcanic ridges; in others, volcanoes are scattered over the top and sides of the ridge (Figs 2, 3).

Deep-towed side-scan sonar surveys of the axial zone (refs 30, 33, 34; D.K.S. and J.R.C., unpublished data) provide high-resolution images of these small volcanoes (Fig. 4). Examples of the different types of morphologies imaged are shown in Fig. 5. Although the resolution of the side-scan sonar is not good enough to discriminate between individual flow morphologies such as sheet flows and pillow basalts, two distinct textures are evident on the images: smooth and hummocky. (The word hummock is used here to describe rounded mounds 50–500 m in diameter and less than 50 m high.) Near-circular volcanoes are common. They include smooth-textured, flat-topped edifices with and without craters, and domes surrounded by brims of lava. Most typically, near-circular volcanoes are built of hummocks which pile up to form cones and domes. In addition to building circular features, hummocks string together to form ridges a few kilometres long.

The different types of small volcanoes pile upon one another and combine with hummocky and smooth-textured flows to form a patchwork that builds the topography of the inner valley floor: axial volcanic ridges and their flanking deeps^{10,13,14,19,30,31}. The relative proportion of each type of volcanic building block varies spatially both locally and between spreading segments. The relationship between variations in the small-scale volcanism and the large-scale topography and ridge axis segmentation is currently unknown. However, the controls on generating small volcanoes must include volume of magma available, depth to the magma body, eruption rate and shape of the plumbing system below the ridge.

Of the spreading segments between the Kane and Atlantis transforms, only one (near 25° 15' N) has no axial volcanic ridge. Instead it has a shallow (~150 m) central graben extending along much of its length. Deep-towed side-scan images indicate that individual small volcanic edifices are scattered across the valley floor, and are identical to those imaged in other spreading segments (D.K.S. and J.R.C., unpublished data). The volcanoes are surrounded by smooth-textured lava flows (see Fig. 5d). Whether these lava flows are sheet or pillow flows can not be determined from available data, however. Such extensive smooth-textured lava flows were not imaged by the deep-towed side-scan sonar in the eight other segments surveyed in the Kane–Atlantis area, suggesting that eruption conditions in this segment are different. Low-relief flows have been correlated with low-viscosity magmas, fast eruption rates and large volumes of magma³⁵, and are commonly found at fast-spreading ridges such as the EPR.

Even though most of the small-scale extrusive volcanism on the inner valley floor is associated with building axial volcanic ridges there is evidence that the eruptions occur throughout the floor^{33,36,37}. Analysis of rock samples from FAMOUS indicated that in some places lava near the edges of the floor are younger

than those found near the centre³⁶ although, in general, rocks from the edges are older. Deep-towed side-scan images of the axial zone between the Kane and Atlantis transforms suggest that volcanism occurs throughout the floor at a frequency that does not allow significant sediment accumulation. The side-scan images show a sharp decrease in echo amplitude outwards across the first major fault scarp of the bounding walls which we relate to a sharp increase in sediment thickness.

Although the sedimentation rate in the Kane–Atlantis area (~20 m Myr⁻¹; ref. 38) would lead one to predict sediment thicknesses of about 8 m at the edge of the valley floor, observations from submersible dives indicate that sediment thicknesses are typically less than 1 m on the inner valley floor³⁸. In the valley walls, however, sediment accumulations of several metres are found³⁸. These observations suggest that volcanism decreases abruptly in the walls of the median valley allowing sediment to build up. Lava must, therefore, cover more or less frequently most parts of the inner valley floor.

Faulting on the inner valley floor. Submersible observations, photographs, and deep-towed side-scan sonar images show that within the inner valley floor the volcanic topography is fractured and fissured, and small-throw normal faults are common^{8,30–32,34,38} (Figs 4 and 5). Overlapping and adjacent small volcanoes sometimes display different degrees of faulting, revealing age relationships. Such an association is useful for deciphering the volcanic stratigraphy, and providing information on the spatial distribution of recent volcanic activity.

New inner-valley-floor bounding faults nucleate within the floor as a result of extensional stresses. Bounding scarps step inwards by one-third to one-half of the width of the inner floor each time a new major fault forms, as crustal blocks that compose the median valley walls are typically several kilometres wide⁷. The morphology of individual fault blocks in the bounding walls suggests that most contain a substantial portion of what is inferred to be an axial volcanic ridge³³. If the identification of these ridges as fossil axial volcanic ridges is correct, the inward step of the bounding scarp often takes it well into or even to the opposite flank of a volcanic ridge.

The structurally defined edge of the inner valley floor does not follow a single fault scarp for the entire length of a segment^{12,29}. Interpretations of side-scan images suggest that individual major bounding faults terminate within the inner valley floor, dying out into a line of minor discontinuous faults sometimes cutting an axial volcanic ridge (D.K.S. and J.R.C., unpublished data). Submersible observations indicate that fissures and small faults are common on the flanks of axial volcanic ridges, and it has been proposed that these are the principal sites at which major bounding scarps nucleate⁸. New bounding faults may step inwards within a segment, or more likely, propagate into the valley floor from a neighbouring and slightly offset segment. Faults grow by propagation along the segment, sometimes splitting an axial volcanic ridge, and isolating crustal blocks which are then incorporated into the bounding walls. Because of this, the width of the inner valley floor, the dimensions of an axial volcanic ridge, and the position of an axial volcanic ridge within the inner floor vary dramatically along a segment. If the bounding fault predominantly steps in from the same side, apparent asymmetric spreading occurs locally. This may be balanced later by asymmetric spreading in the opposite sense.

Magma supply, storage and plumbing

Despite all the recent observations of bathymetry, gravity anomalies, seismic structure, magnetic structure and basalt geochemistry we still know very little about how magma is supplied from the mantle to the crust. Mantle upwelling is probably enhanced at segment centres where sea-floor depths are generally shallowest^{4,11,24,28}, so that melt supply is greater at the centre than at the ends giving, in general, thicker crust at segment centres^{27,28}. Melt may accumulate within the crust near the centre of segments and then migrate laterally. Alternatively, the move-

ment of magma may be predominantly vertical along the entire segment, with a greater supply at segment centres reflecting more vigorous upwelling there.

From geophysical evidence it is inferred that a melt lens exists beneath some parts of the axis of the EPR, and can be continuous for several tens of kilometres along the axis^{39–41}. On the other hand, geophysical studies in several places at the MAR have failed to identify a melt lens^{26,42–47} although a few studies suggest the presence of low-velocity zones in the lower crust where there is significant attenuation of compressional wave energy. A seismic refraction investigation in the MARK area indicated a region of lower velocities and presumably higher temperatures 5–6 km below an axial volcanic ridge (Snakepit Ridge)²⁶; similarly, results from a microearthquake survey in the TAG area can be explained by the existence of a low-velocity volume 4–6 km below a bathymetric high⁴⁷. Both of these low-velocity regions are interpreted to be the partially cooled remnants of a recent injection of magma, rather than melt (because they support shear-wave propagation and limited faulting), which suggests that crustal magma chambers are transient features at slow-spreading ridges.

Magma storage in the crust. For magma to ascend into the crust it must be positively buoyant. This occurs when the density of the magma is less than that of the surrounding rock. Assuming that pathways exist for magma to rise along, it will halt in its ascent when it reaches a level of neutral buoyancy where the density of the magma is the same as the surrounding materials. It may also halt before the neutral buoyancy level, if it reaches an impenetrable horizon. Two important scientific questions concern the size of the magma bodies at the MAR, and what controls the depth of magma storage in the crust.

Early models placed a large shallow crustal magma chamber below the MAR valley floor³⁶. Nisbet and Fowler⁴⁸, however, after a re-evaluation of existing seismic and petrological constraints proposed that magma is supplied to the MAR in small batches that rise to the base of the crust and crystallize there to form small plutons. Magma then rises through cracks from these depths to feed eruptions on the sea floor. We recently suggested that rather than stopping at the base of the crust, magma rises buoyantly within the crust to the sharp rheological transition at the base of the brittle lid, where magma is trapped to form small bodies¹⁴. Hydrothermal circulation within the upper part of the ocean crust at the spreading axis leads to the formation of a strong brittle lid above the plastic asthenosphere^{49,50}. Other workers have proposed that magma accumulates at a freezing horizon where heat input from magma injection is balanced by heat loss from conductive and hydrothermal cooling⁵¹.

To place constraints on the depth at which the melt is trapped at the MAR we used magma buoyancy arguments⁵² and the heights of small flat-topped volcanoes on the inner valley floor¹⁴. The buoyancy argument is that the magmatic pressure in the magma chamber due to the vertical column of magma should equal the lithostatic and hydrostatic pressure exerted on the magma chamber by the adjacent column of rock and water. The flat-topped cone morphology of small isolated volcanoes is thought to be caused by the cone reaching the magmatic limit of eruption⁵³. Thus, a volcano grows as a peaked cone until it reaches the limiting height, and then builds outwards at the summit, creating a flat top. The height of a flat-topped volcano above the surrounding sea floor, therefore, should be related to the depth of the magma source and the densities of the magma and surrounding materials. We calculated that volcano heights on the inner valley floor of the MAR in the Kane–Atlantis area correspond to magma source depths of 1–7 km below the sea floor¹⁴.

The hypothesis that magma is halted at a sharp rheological boundary such as the base of the crust or the base of the brittle lid is challenged by the alternative view that magma rises to a level of neutral buoyancy^{54,55}. This may be the case in subaerial volcanoes such as Hawaii^{54,56}. In the oceanic crust, a variety of

lines of evidence (logging of deep-ocean drill holes^{57,58}, seismic structure studies⁴⁰ and ophiolite structure^{59,60}) imply that the level of neutral buoyancy (100–400 m below the sea floor)⁶¹ lies well above the level at which magma is stored within the crust. The question of where rising magma stops is currently a topic of much debate.

Plumbing beneath the ridge. Despite the fact that we cannot specifically state why rising magma halts its ascent in the crust, we do know that from the zone of accumulation, magma rises to feed the extrusives. On a large scale, the extrusives are organized into axial volcanic ridges, which may indicate some segment-scale structure in magmatic plumbing. The fact that axial volcanic ridges are composed of small volcanoes also has important implications for the plumbing system beneath the axis. Small ridges composed of hummocks are clearly formed along linear cracks (Fig. 5d). The individual hummocks may represent a focusing of an eruption along an original fissure as observed in Hawaii⁶², and Iceland⁶³. An illustrative model⁶⁴ of the eruptive process shows that when the rates of eruption are high, flow is fast enough to discharge along the whole length of the fissure. When the rates of eruption are low, however, cooling increases the viscosity of the flowing magma and the original fissure flow breaks down into vertical fingers. It can be inferred from the multitude of small hummocky ridges that, in general, eruption rates are not fast enough to sustain the eruption along the entire length of a fissure.

The feeder systems below circular features such as cones and domes are harder to deduce. They also may represent a focusing of an eruption along an original fissure. However, some of the small volcanoes show no sign of an associated fissure. For example, in the Kane–Atlantis area, deep-tow side-scan images show a hummocky cone (120 m in relief from multibeam bathymetry) surrounded by a broad smooth low-relief lava flow with no feeding fissure in evidence. Whether this feature erupted onto the smooth flow from a pipe or a short fissure now entirely covered by its flanks, or whether it erupted from a longer fissure subsequently covered by the smooth flow can not be determined from the morphology alone. The dominance of fissures in extensional environments such as ophiolites^{65,68}, and the observation of what are thought to be sheeted dikes in the valley walls³⁸ and fracture zones⁶⁹ at the MAR, support the conclusion that most features are formed by eruptions through fissures. Feeders probably take on a range of cross-sectional shapes (circular, elliptical, linear) and linear dimension.

Of course, several volcanic edifices may be produced from a single eruptive vent such as is commonly seen in Hawaii. For example, blocked feeder tubes often build 'rootless' volcanoes called tumuli⁶². Serocki Volcano, a 60-m flat-topped edifice on the inner valley floor in the MAR area, is thought to be a megatumulus⁷⁰ which built when a flow issuing from an adjacent volcano was blocked (W. B. Bryan and S. Humphris, personal communication). In addition, lava may spill over from an original eruptive vent and build a perched lava pond such as now exists at Mauna Ulu on the east rift of Kilauea Volcano on Hawaii. Perched lava ponds equivalent in dimensions to neighbouring edifices could easily be misidentified as individual flat-topped volcanoes. It is clear that additional constraints (as well as the morphology) are necessary to sort out the complex plumbing system beneath the MAR.

Structure of the crust. Two of the most important controls at a slow-spreading centre that yield abundant small volcanoes must be the low magma flux and the great depth of hydrothermal circulation. Low magma flux will produce small, isolated magma bodies^{48,71} and a thicker brittle lid⁴⁹. Microearthquake studies at the MAR^{44,47} indicate that brittle failure commonly occurs to depths greater than 4.5 km. From these results it is inferred that faulting occurs to these depths, possibly providing efficient pathways for hydrothermal circulation and cooling⁴⁷. If hydrothermal cooling penetrates to these depths it must play a role in limiting the size and shape of crustal magma bodies.

An early model for the structure of the oceanic crust at a spreading ridge is shown in Fig. 6a (ref. 72). It contains a single large magma chamber; a sheeted dike complex lies above the magma chamber and feeds the extrusives. A model proposed for the FAMOUS segment is displayed in Fig. 6b (ref. 36); it also contains a large ridge-axis magma chamber. If, however, as evidence now implies, the MAR is fed by small ephemeral magma pockets the structure of the lower oceanic crust will be substantially different from that shown in Fig. 6a and b. Instead, the lower crust will be made up of a multitude of small plutonic bodies, each linked to (a) surface volcanic feature(s) by a feeder dike or pipe, and each having cooled separately (Fig. 6c and d). The resulting structure is close to that shown by Moores and Vine⁶⁷ for the plutonic part of the Troodos ophiolite in Cyprus where they suggest that there have been multiple intrusions and extrusions of small volumes of melt.

Furthermore, feeding by fissures or pipes has consequences for the transition between extrusives and plutonics. Where flows are principally fed by fissures, the density of dikes leads to a sheeted dike complex, especially if the intrusion of dikes is confined to a narrow zone at the spreading axis^{73,74}. Where extrusives are fed primarily by pipes, there may be no zone composed entirely of feeders linking plutonics to extrusives. Instead extrusives may persist, although with an increasing density of feeder pipes or dikes, down to the top of the plutonics³⁷.

The structure of the oceanic crust at the MAR is thought to be highly variable along the length of a spreading segment^{56,47,75}. In the centre of a segment, above a zone of focused mantle upwelling^{27,28}, enhanced temperatures may result in relatively large magma bodies¹². Near the ends of segments, melt production is probably lower, and the oceanic crust may be built by smaller magma pockets. How magma is supplied to an entire spreading segment is unknown, and remains to be answered in the future.

Building the crust

Since the FAMOUS project it has been argued that MAR magmatic and tectonic activity is cyclic^{6,10,76,77}. Episodes dominated by volcanic construction are thought to alternate with amagmatic, tectonically dominated extension. It is clear from their morphology that individual ridge segments are characterized by different ratios of magmatic/tectonic activity: one segment will contain a robust axial volcanic ridge that covers most of the inner valley floor while another contains a smaller axial volcanic ridge dissected by faults (Fig. 2). At the same time, it is clear that these differences extend to smaller scales: an individual spreading segment also exhibits along-axis variations in morphological characteristics. Within a spreading segment an axial volcanic ridge may change from a robust feature unaffected by faulting, to two or more subparallel and adjacent ridges, to one that is being cut by faults, to a discontinuous ridge (Fig. 2). There are also intrasegment variations in major bounding fault length, spacing and throw^{29,78}. Such evidence argues against segments acting as single units and going through cycles. Rather, we infer that processes at the MAR are three-dimensional over many scales.

Figure 7 illustrates a general model for building the oceanic crust at the northern MAR that allows for these along-axis variations. Beneath the ridge axis, discrete small batches of magma rise buoyantly within the shallow crust and are trapped at a sharp rheological boundary. The boundary varies in position through time between about 2–5 km below the sea floor in response to the balance between local magma supply and hydrothermal cooling. We visualize the magma bodies as flat lenses or sills which, at the slow spreading rate of the northern MAR, will most likely solidify before the next is emplaced, and the lower crust will be made of a multitude of small plutons. Dikes and pipes link the magma bodies to the sea floor of the inner valley floor where eruptions construct small volcanoes with diverse morphologies. Eruptions occur throughout the length

and width of the valley floor, though more frequently towards its centre. Eruptive events focus to build axial volcanic ridges composed of piled-up small volcanoes and flows. The dimensions and character of the axial volcanic ridge varies along a segment in response to variations in the supply of melt, eruption rates, and the plumbing system.

The newly constructed sea-floor is modified by tectonic processes. As crust is formed and the sea floor spreads, the bounding scarps of the inner valley floor step inward. A new fault initiates on the inner valley floor or propagates into the segment from an adjacent and slightly offset segment. This style of faulting means that at any one time there is no single fault that represents the structural edge of the inner valley floor. The magnitude of the inward step of the fault is typically large enough to take the fault into an axial volcanic ridge, or even to its further side. Because of the complexity of faulting and its interaction with volcanism, an axial volcanic ridge may be strongly modified by faulting in one part of a segment, and relatively unaffected by faulting in another part, producing marked along-axis morphological variations.

The inward step of the bounding fault which often cuts an axial volcanic ridge will lead to the incorporation of portions of the volcanic ridge in the median valley walls and eventually the crestal mountains. At the same time, in that section of the spreading segment, the principal axis of accretion in the now narrower inner valley floor moves laterally by a few kilometres, and a new axial volcanic ridge is constructed. The remaining section of the older axial volcanic ridge (now asymmetrically located on the inner floor) may continue to be volcanically active. The lateral shift of the primary locus of extrusive activity produces local apparent asymmetric spreading. In the long run, however, because the process can occur from either side of the axis, symmetric spreading normally occurs.

The three-dimensional interplay of volcanism and tectonism produces both morphological variation within a segment, and contrasting morphologies between segments. Some segments contain large, robust axial volcanic ridges, others contain smaller ridges cut by myriad fissures and small faults; some segments have narrow inner valley floors, and others wide ones. The distinctive variability of the topography produced at the northern MAR is quite different from that observed at the fast-spreading EPR¹⁵. This difference can be partly attributed to the existence of a steady-state melt lens along large sections of the EPR, which creates a structure that is basically two-dimensional, at least compared to the fundamentally three-dimensional structure of the MAR.

The newly obtained images of the topography of the axial zone of the northern MAR have clarified some of the previous difficulties in interpreting the process of crustal construction. The evidence from high-resolution morphology of long sections of the axial zone has, though, emphasized the fundamental complexity of the crustal construction process, and the need both for new insights and new critically directed observations. □

D. K. Smith is at the Department of Geology and Geophysics, Woods Hole Oceanographic Institute, Woods Hole, Massachusetts 02543, USA. J. R. Cann is at the Department of Earth Sciences, University of Leeds, Leeds LS2 9JT, UK.

1. Menard, H. W. *Science* **157**, 923–924 (1967).
2. van Andel, T. J. H. & Bowin, C. O. *J. geophys. Res.* **73**, 1279–1298 (1968).
3. Lonsdale, P. *Mar. geophys. Res.* **3**, 251–293 (1977).
4. Macdonald, K. C. in *The Geology of North America* (ed. Winterer, E. L.) 93–111 (Geol. Soc. Am., Boulder, 1989).
5. Macdonald, K. C. & Fox, P. J. *Earth planet. Sci. Lett.* **88**, 119–131 (1988).
6. Macdonald, K. C. in *The Geology of North America* (eds Vogt, P. R. & Tucholke, B. E.) 51–68 (Geol. Soc. Am., Boulder, 1986).
7. Macdonald, K. C. *Geol. Soc. Am. Bull.* **88**, 541–555 (1977).
8. Ballard, R. D. & van Andel, T. J. *Geol. Soc. Am. Bull.* **88**, 507–530 (1977).
9. Ramberg, I. B. & van Andel, T. J. *Geol. Soc. Am. Bull.* **88**, 577–586 (1977).
10. Karson, J. A. et al. *Nature* **328**, 681–685 (1987).
11. Sempere, J.-C., Purdy, G. M. & Schouten, H. *Nature* **344**, 427–431 (1990).
12. Sempere, J.-C., Lin, J., Brown, H. S., Schouten, H. & Purdy, G. M. *Mar. geophys. Res.* **15**, 153–200 (1993).

13. Smith, D. K. & Cann, J. R. *Nature* **348**, 152–155 (1990).
14. Smith, D. K. & Cann, J. R. *J. geophys. Res.* **97**, 1645–1658 (1992).
15. Macdonald, K. C. et al. *Nature* **335**, 217–225 (1988).
16. Heezen, B. C. & Tharp, M. *Geol. Soc. Am. Map* (Geol. Soc. Am., Boulder, 1957).
17. Aumento, F., Loncarevic, B. D. & Ross, D. I. *Phil. Trans. R. Soc.* **268**, 623–650 (1971).
18. Heitzler, J. R. & van Andel, T. J. *Geol. Soc. Am. Bull.* **88**, 481–487 (1977).
19. Crane, K. & Ballard, R. D. *J. geophys. Res.* **86**, 5112–5124 (1981).
20. Scott, R. B., Rona, P. A., McGregor, B. A. & Scott, M. R. *Nature* **251**, 301–302 (1974).
21. Phillips, J. D. & Fleming, H. S. *Geol. Soc. Am. Maps MC-19* (Geol. Soc. Am., Boulder, 1978).
22. Purdy, G. M., Sempere, J.-C., Schouten, H., Dubois, D. L. & Goldsmith, R. *Mar. geophys. Res.* **12**, 247–252 (1990).
23. Ramberg, I. B., Gray, D. F. & Reynolds, R. G. H. *Geol. Soc. Am. Bull.* **88**, 609–620 (1977).
24. Crane, K. *Earth planet. Sci. Lett.* **72**, 405–414 (1985).
25. Schouten, H., Klitgord, K. D. & Whitehead, J. A. *Nature* **317**, 225–229 (1985).
26. Purdy, G. M. & Detrick, R. S. *J. geophys. Res.* **91**, 3739–3762 (1986).
27. Kuo, B. Y. & Forsyth, D. W. *Mar. geophys. Res.* **10**, 205–232 (1988).
28. Lin, J., Purdy, G. M., Schouten, H., Sempere, J.-C. & Zervas, C. *Nature* **344**, 627–632 (1990).
29. Shaw, P. R. *Nature* **358**, 490–493 (1992).
30. Kong, L. S., Detrick, R. S., Fox, P. J., Mayer, L. A. & Ryan, W. B. F. *Mar. geophys. Res.* **10**, 59–90 (1988).
31. Brown, J. R. & Karson, J. A. *Mar. geophys. Res.* **10**, 109–138 (1988).
32. Gente, P., Mevel, C., Auzende, J. M., Karson, J. A. & Fouquet, Y. *Tectonophysics* **190**, 1–29 (1991).
33. Macdonald, K. C. & Luyendyk, B. P. *Geol. Soc. Am. Bull.* **88**, 621–636 (1977).
34. Luyendyk, B. P. & Macdonald, K. C. *Geol. Soc. Am. Bull.* **88**, 648–663 (1977).
35. Bonatti, E. & Harrison, C. G. A. *J. geophys. Res.* **93**, 2967–2980 (1988).
36. Bryan, W. B. & Moore, J. G. *Geol. Soc. Am. Bull.* **88**, 556–570 (1977).
37. Atwater, T. in *Deep Drilling Results in the Atlantic Ocean: Ocean Crust Vol. 2* (eds Talwani, M., Harrison, C. G. & Hayes, D. E.) 33–42 (Am. geophys. Un., Washington DC, 1979).
38. Zonenshain, L. P. et al. *Tectonophysics* **159**, 1–23 (1989).
39. Detrick, R. S. et al. *Nature* **326**, 35–41 (1987).
40. Herron, T. J., Ludwig, W. J., Stoffa, P. L., Kan, T. K. & Buhl, P. J. *geophys. Res.* **83**, 798–804 (1978).
41. Sinton, J. M. & Detrick, R. S. *Mar. geophys. Res.* **97**, 197–216 (1992).
42. Whitmarsh, R. B. *Nature* **246**, 297–299 (1973).
43. Fowler, C. M. & Keen, C. E. *Geophys. J. R. astr. Soc.* **56**, 219–226 (1979).
44. Toomey, D. R., Solomon, S. C. & Purdy, G. M. *J. geophys. Res.* **93**, 9093–9112 (1988).
45. Huang, P. Y. & Solomon, S. C. *J. geophys. Res.* **93**, 13445–13477 (1988).
46. Detrick, R. S., Mutter, J. C., Buhl, P. & Kim, I. I. *Nature* **347**, 61–64 (1990).
47. Kong, L. S., Solomon, S. C. & Purdy, G. M. *J. geophys. Res.* **97**, 1659–1685 (1992).
48. Nisbet, E. G. & Fowler, C. M. R. *Geophys. J. R. astr. Soc.* **54**, 631–660 (1978).
49. Harper, G. *Tectonics* **4**, 395–409 (1985).
50. Chen, Y. & Jason Morgan, W. J. *geophys. Res.* **95**, 17583–17604 (1990).
51. Phipps-Morgan, J. & Chen, Y. J. *J. geophys. Res.* **98**, 6283–6297 (1993).
52. Daly, R. A. *Igneous Rocks and the Depths of the Earth* (McGraw-Hill, New York, 1933).
53. Barone, A. M. & Ryan, W. B. F. *J. geophys. Res.* **95**, 10801–10827 (1990).
54. Ryan, M. P. in *Magmatic Processes: Physicochemical Principles* (ed. Mysen, B. O.) 259–287 (Geochem. Soc. spec. Publ., University Park, Pennsylvania, 1987).
55. Wilson, L., Head, H. W. & Parfitt, E. A. *Geophys. Res. Lett.* **19**, 1395–1398 (1992).
56. Tilling, R. I. & Dvorak, J. J. *Nature* **363**, 125–132 (1993).
57. Becker, K. *Init. Rep. DSDP Leg 83*, 419–427 (1985).
58. Becker, K., Langseth, M. G., Von Herzen, R. P., Anderson, R. N. & Hobart, M. A. *Init. Rep. DSDP Leg 83*, 405–418 (1985).
59. Gass, I. G. & Smewing, J. D. in *The Sea 7: The Oceanic Lithosphere* (ed. Emiliani, C.) 339–362 (Wiley, New York, 1981).
60. Harper, G. *Geol. Soc. Am. Bull.* **95**, 1009–1026 (1984).
61. Hoof, E. E. & Detrick, R. S. *Geophys. Res. Lett.* **20**, 423–426 (1993).
62. Macdonald, G. A. & Abbott, A. T. *Volcanoes in the Sea* (Univ. Hawaii Press, Honolulu, 1970).
63. Walker, G. P. L. *Leicester Lit. Phil. Soc. Trans.* **51**, 25–40 (1964).
64. Whitehead, J. A. & Helfrich, K. R. *J. geophys. Res.* **96**, 4145–4155 (1991).
65. Gass, I. G. *Geol. Surv. Cyprus Mem.* **4**, 1–116 (1960).
66. Gass, I. G. *Nature* **220**, 39–42 (1968).
67. Moores, E. M. & Vine, F. J. *Phil. Trans. R. Soc. A268*, 443–466 (1971).
68. Wilson, R. A. M. *Geol. Surv. Cyprus Mem.* **1**, 1–135 (1959).
69. Auzende, J.-M. et al. *Nature* **337**, 726–729 (1989).
70. Humphris, S. E., Bryan, W. B., Thompson, G. & Autio, L. K. *Proc. ODP Sci. Res.* **106/109**, 67–84 (1990).
71. Perry, F. V., Baldrige, W. S., DePaolo, D. J. & Shafiqullah, M. *J. geophys. Res.* **95**, 19327–19348 (1990).
72. Cann, J. R. *Geophys. J. R. astr. Soc.* **39**, 169–187 (1974).
73. Atwater, T. & Mudie, J. D. *J. geophys. Res.* **78**, 8665–8686 (1973).
74. Kidd, R. G. W. *Geophys. J. R. astr. Soc.* **50**, 149–183 (1977).
75. Solomon, S. C. & Toomey, D. R. A. *Rev. Earth planet. Sci.* **20**, 329–364 (1992).
76. Gentre, P. et al. *C. R. Seanc. Acad. Sci., Paris* **308**, 1781–1788 (1989).
77. Eberhart, G. L., Rona, P. A. & Honnorez, J. *Mar. geophys. Res.* **10**, 233–259 (1988).
78. Shaw, P. R. & Lin, J. *J. geophys. Res.* (in the press).

ACKNOWLEDGEMENTS. We thank the master, crew and scientific party of RV *Charles Darwin*, and the Institute of Oceanographic Sciences, Deacon Lab., for providing and supporting TOBI at sea. Funding was from the NERC and the US Office of Naval Research.

A search for life on Earth from the Galileo spacecraft

Carl Sagan^{*}, W. Reid Thompson^{*}, Robert Carlson[†], Donald Gurnett[‡] & Charles Hord[§]

^{*} Laboratory for Planetary Studies, Cornell University, Ithaca, New York 14853, USA

[†] Atmospheric and Cometary Sciences Section, Jet Propulsion Laboratory, Pasadena, California 91109, USA

[‡] Department of Physics and Astronomy, University of Iowa, Iowa City, Iowa 52242-1479, USA

[§] Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder, Colorado 80309, USA

In its December 1990 fly-by of Earth, the Galileo spacecraft found evidence of abundant gaseous oxygen, a widely distributed surface pigment with a sharp absorption edge in the red part of the visible spectrum, and atmospheric methane in extreme thermodynamic disequilibrium; together, these are strongly suggestive of life on Earth. Moreover, the presence of narrow-band, pulsed, amplitude-modulated radio transmission seems uniquely attributable to intelligence. These observations constitute a control experiment for the search for extraterrestrial life by modern interplanetary spacecraft.

At ranges varying from ~100 km to ~100,000 km, spacecraft have now flown by more than 60 planets, satellites, comets and asteroids. They have been equipped variously with imaging systems, photometric and spectrometric instruments extending from ultraviolet to kilometre wavelengths, magnetometers and charged-particle detectors. In none of these encounters has compelling, or even strongly suggestive, evidence for extraterrestrial life been found. For the Moon, Venus and Mars, orbiter and lander observations confirm the conclusion from fly-by spacecraft. Still, extraterrestrial life, if it exists, might be quite unlike the forms of life

with which we are familiar, or present only marginally. The most elementary test of these techniques—the detection of life on Earth by such an instrumented fly-by spacecraft—had, until recently, never been attempted.

Galileo is a single-launch Jupiter orbiter and entry probe currently in interplanetary space and scheduled to arrive in the Jupiter system in December 1995. It could not be sent directly to Jupiter; instead, the mission incorporated two close gravitational assists at the Earth and one at Venus. This greatly lengthened the transit time, but it also permitted close observations of the Earth. The

Galileo instruments were not designed for an Earth encounter mission, so an appropriate control experiment has fortuitously been arranged: a search for life on Earth with a typical modern planetary probe.

Closest approach to Earth in the 8 December 1990 encounter was 960 km over the Caribbean Sea. The Earth was approached from its night side, so all data in reflected light were acquired post-encounter. Evidence relevant to the search for life was obtained with the near-infrared mapping spectrometer (NIMS), the ultra-violet spectrometer (UVS), the solid-state imaging system (SSI), and the plasma wave spectrometer (PWS). These instruments are described in refs 1–4, respectively. Because of the high encounter velocity, most NIMS and SSI data were obtained 2 hours to 3.5 days after closest approach.

In what follows, we do not assume properties of life otherwise known on Earth, but instead attempt to derive our conclusions from Galileo data and first principles alone. We compare these conclusions with ground-based and low-altitude measurements that we describe, collectively, as the 'ground-truth Earth'. A necessary but not sufficient condition for the presence of life is a marked departure from thermodynamic equilibrium^{5–8}. Once candidate disequilibria are identified, alternative explanations must be eliminated. Life is the hypothesis of last resort. As an analogy, mountains are in mechanical disequilibrium, given the Earth's erosional environment, but orogenic processes build mountains faster than wind, water and plate tectonics destroy them. Ozone is in substantial thermodynamic disequilibrium in the Earth's atmosphere, but this is driven by solar ultraviolet (UV) photochemistry. Bursts of high-brightness-temperature kilometre-wave radiation occur in the Earth's auroral zones, but are pumped by the interaction of the solar wind with the Earth's magnetosphere. No biological explanations of these disequilibrium phenomena need to be sought. As we will show, however, Galileo found such profound departures from equilibrium that the presence of life seems the most probable cause.

Chemistry

NIMS spectral and radiometric measurements indicate the presence of water in several forms. Spectra of Antarctica⁹ show distinctive features due to condensed water, which are broadened and shifted to longer wavelengths compared to gas-phase

features. The radiometric temperatures ($\sim 30^\circ\text{C}$) indicate ice, and analysis of the spectra⁹ give snow grain sizes of 50–200 μm . This ice and snow cover occurs on continental scale around the South Pole.

At higher latitudes are extensive areas with higher, nearly uniform temperatures just at, or slightly above, the melting point of water. In particular, the radiometric temperatures measured at 5 μm , uncorrected for atmospheric absorption, are $\sim -3^\circ\text{C}$ at high southern latitudes, increasing to 8–18 $^\circ\text{C}$ at midlatitudes. The average 1 μm albedo of these extensive areas is $\sim 4\%$, much smaller than the albedos of snow, clouds and rocky surfaces, but consistent with the low diffuse reflectance of dielectric liquid surfaces, including water. In many of the NIMS images, greatly enhanced specular reflection is observed¹⁰—implying the existence of large areas that are macroscopically smooth and homogeneous (that is, not granular) and most easily explained by the presence of liquid surfaces of oceanic dimensions. Evidence of gas-phase H_2O is found over the entire planet. Representative NIMS infrared spectra in the 0.7–1.0 μm range, and in the 2.4–5.2 μm range, are shown in Fig. 1. They were obtained over a fairly clear area in the eastern Pacific, north of Borneo. Analysis¹⁰ of the vibration–rotation bands of water vapour gives typical abundances of $\sim 1,000$ parts per million (p.p.m.) or $\sim 0.6\text{ g cm}^{-2}$. The observed high humidities, found over most of the planet, along with the preceding discussion, imply that the oceans are composed of liquid water.

Spectral data aside, from the albedo and heliocentric distance of the Earth alone it follows that the equilibrium temperature of the planet is only 20 $^\circ\text{C}$ or so below the freezing point of water—so that even a modest greenhouse effect would bring temperatures high enough that water, a cosmically abundant molecule, could exist in all three of its phases. Abundant surface liquid water is seen nowhere else in the contemporary Solar System. The dielectric constant, solvation properties, heat capacity, and temperature range of the liquid state are among the nonparochial reasons that water seems an ideal medium for life¹¹.

Figure 1 is notable for the presence of the A band of molecular oxygen at 0.76 μm . This transition is spin-forbidden, and the strength of the feature indicates $\sim 200\text{ g cm}^{-2}$ of O_2 (Table 1). So large an abundance of O_2 is also unique among all the worlds

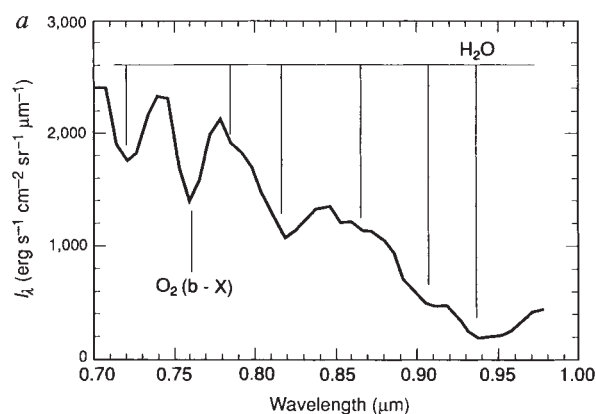
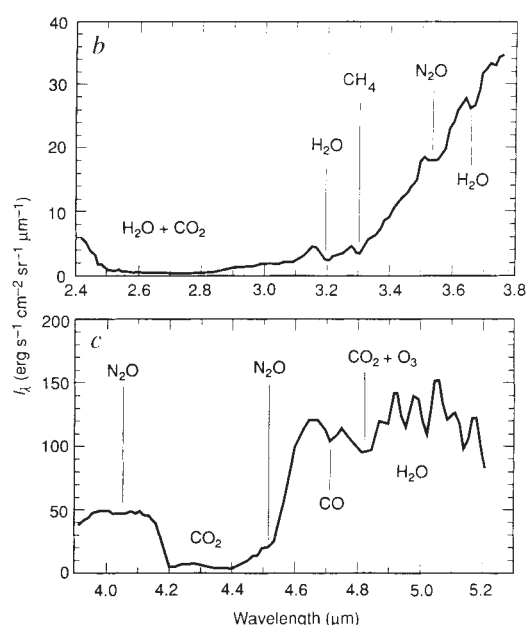


FIG. 1 a, Galileo long-wavelength-visible and near-infrared spectra of the Earth over a relatively cloud-free region of the Pacific Ocean, north of Borneo. The incidence and emission angles are 77° and 57° respectively. The $(b' - X)$ O_2 band at 0.76 μm is evident, along with a number of H_2O features. Using several cloud-free regions of varying airmass, we estimate an O_2 vertical column density of $1.5\text{ km-atmagat} \pm 25\%$. b and c, Infrared spectra of the Earth in the 2.4–5.2 μm region. The strong ν_3 CO_2 band is seen at the 4.3 μm , and water vapour bands are found, but not indicated, in the 3.0 μm region. The ν_3 band of nitrous oxide, N_2O , is apparent at the edge of the CO_2 band near 4.5 μm , and N_2O combination bands are also seen near 4.0 μm . The



methane (0010) vibrational transition is evident at 3.31 μm . A crude estimate¹⁰ of the CH_4 and N_2O column abundances is, for both species, of the order of 1 cm-atmagat ($\equiv 1\text{ cm path at STP}$).

TABLE 1 Constituents of the Earth's atmosphere (volume mixing ratios)

Molecule	Standard abundance (ground-truth Earth)	Galileo value*	Thermodynamic equilibrium value Estimate 1† Estimate 2‡
N ₂	0.78		0.78
O ₂	0.21	0.19 ± 0.05	0.21§
H ₂ O	0.03–0.001	0.01–0.001	0.03–0.001
Ar	9 × 10 ⁻³		9 × 10 ⁻³
CO ₂	3.5 × 10 ⁻⁴	5 ± 2.5 × 10 ⁻⁴	3.5 × 10 ⁻⁴
CH ₄	1.6 × 10 ⁻⁶	3 ± 1.5 × 10 ⁻⁶	< 10 ⁻³⁵ 10 ⁻¹⁴⁵
N ₂ O	3 × 10 ⁻⁷	~10 ⁻⁶	2 × 10 ⁻²⁰ 2 × 10 ⁻¹⁹
O ₃	10 ⁻⁷ –10 ⁻⁸	> 10 ⁻⁸	6 × 10 ⁻³² 3 × 10 ⁻³⁰

* Galileo values for O₂, CH₄ and N₂O from NIMS data; O₃ estimate from UVS data.

† From ref. 16 (P, 1 bar; T, 280 K).

‡ From ref. 17 (P, 1 bar; T, 298 K).

§ The observed value; it is in thermodynamic equilibrium only if the under-oxidized state of the Earth's crust is neglected.

in the Solar System. Oxygen can be generated by the UV photodissociation of water and the subsequent Jeans escape of H to space. But can the accumulation of so much O₂ over geological time be understood?

Certainly, Venus and Mars—where atmospheric water vapour

is being UV photodissociated—display very low O₂ abundances, <1 and ~10⁻³ p.p.m., respectively^{12,13}, despite large quantities of water lost through photodissociation and atmospheric escape of H. But a real understanding of the Earth's steady-state O₂ abundance requires at the least knowledge of the oxidation state of the surface, surface erosion rates, and temperatures at the tropopause, the mesopause and the exobase. Galileo did not provide this data set. Even if it had, we could not be confident that present circumstances are typical of Earth history. An upper bound on the UV generation rate of O₂ is set by the photon-limited water photolysis rate, ~10¹³ cm⁻² s⁻¹ on the ground-truth Earth¹⁴—which yields the present O₂ abundance in <10⁵ yr. If instead we use the ground-truth present H escape flux¹⁴ (which is H₂O diffusion-limited) and assume that all the H atoms are H₂O-derived, the present O₂ abundance would require several times the age of the Earth to accumulate. If there is substantial oxidation of the crust, these timescales will be longer. But the lack of impact craters in Galileo imagery, and pervasive wind and water on Earth suggest continuing exposure of fresh, oxidizable regolith. Accordingly, oxidation of the Earth's crust should be more extensive than on Venus and Mars, and yet these planets have much less atmospheric O₂. Galileo's observations of O₂ thus at least raise our suspicions about the presence of life. If detailed modelling showed solar UV to be insufficient, a process seems needed whereby the much more abundant, but much less energetic visible light photons are used in series for

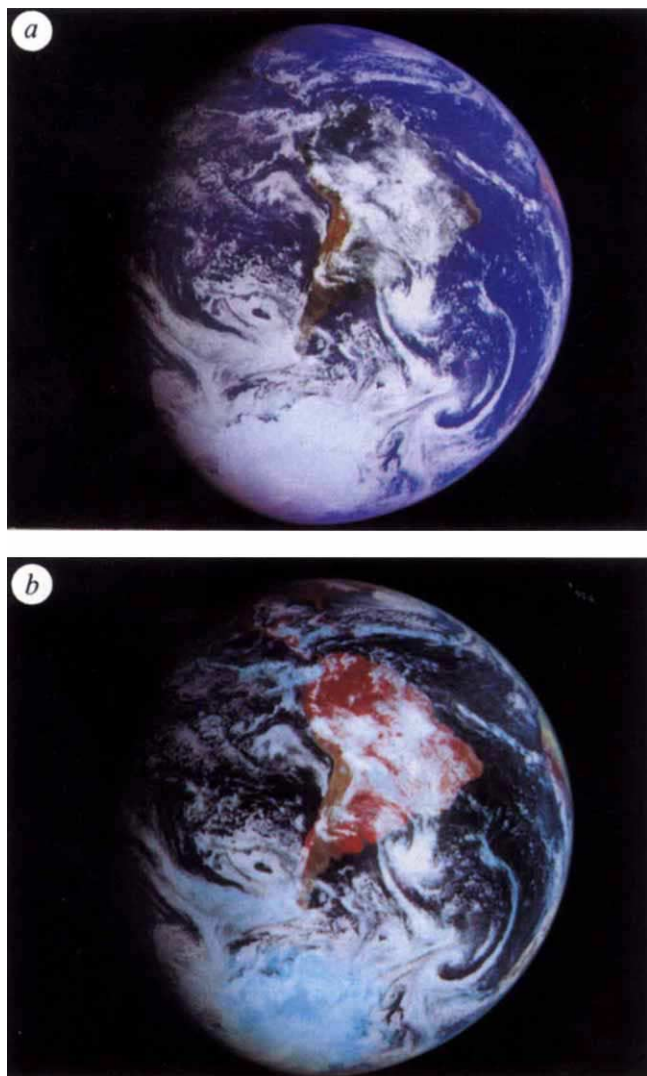


FIG. 2 Three colour-composite versions of a Galileo view of Earth centred over 72° W, 34° S. The SSI bands are identified by the following mnemonics and mean wavelengths in micrometres: VIO (0.407), GRN (0.558), RED (0.670), 0.73 (0.734/N), 0.76 (0.760/N), 0.89 (0.887/N), 1.0 (0.984). N indicates a narrow-band filter (the others are broad-band), and effective wavelengths and radiometric calibration for data presented here are for the SSI camera with the clear lens cap on, which was the case for the encounters described here. Of these seven bands, six (all except 0.89) were used for global imaging of the Earth at 20–30 km resolution. (Galileo also imaged Australia and Antarctica at higher resolution (~ one to a few km per pixel), but these imaging sequences did not use bands beyond 0.73 µm). a, Approximately natural-colour view constructed from (RED, GRN, VIO) filters; b, colour-composite produced from mapping the (1.0 µm, RED, GRN) bands to red, green and blue; c, colour-composite produced from closely spaced (0.76 µm, 0.73 µm, RED) bands, which uniquely separate rock and soil (grey tones) from areas that have unusual, very steep spectral slopes near 0.7 µm, possibly indicative of photosynthetic pigments in autotrophic terrestrial life. Areas whose spectra are shown in Fig. 3 are marked with A, B, C on this composite.

H₂O photodissociation (at least two visible photons would be needed per photodissociation event). Apparently, only biological systems can accomplish this.

In Fig. 1b and c, the presence of carbon dioxide, methane, nitrous oxide and ozone are indicated. All are greenhouse gases; together with water vapour, they mainly account for the discrepancy between the equilibrium and measured radiometric temperatures of the Earth's surface.

From straightforward photochemical theory it follows¹⁵ that one consequence of the high oxygen abundance is a stratospheric ozone layer opaque in the middle UV. Even a quick look at the Galileo UVS raw data shows the strong Hartley bands of O₃ at wavelengths greater than 0.21 μ m in the spectrum of the Earth but not of Venus. (O₃ absorption is also found in NIMS infrared spectra.) The presence of disequilibrium O₃ is not in itself a sign of life, because UV photochemistry is a disequilibrium process, but the photochemical abundance of O₃ is related approximately logarithmically to the O₂ abundance, so the observed quantity of O₃ does imply a substantial abundance of O₂. Therefore a train of argument may exist¹⁵ from abundant O₃ to abundant O₂ to life. The Hartley bands provide substantial optical depth at wavelengths less than $\sim 0.3 \mu$ m, suggesting that in addition to the oceans (where life could be protected at depth from UV radiation), life on the land is also possible provided structural molecular bond strengths exceed $\sim 50 \text{ kcal mol}^{-1}$. In this category fall a large number of organic functional groups, including C-C (70–130 kcal mol^{-1}), C-H (80–110), C-O (~ 90) and C-N (80–100). Corresponding multiple bonds are even stronger. Of course this is not, by itself, an argument for organic biochemistry on Earth.

The agreement of ground-truth atmospheric abundances with those determined by Galileo is reasonably good (Table 1). Also shown in Table 1 are two independent estimates^{16,17}, made 25 years apart, of the thermodynamic equilibrium abundances expected of minor constituents in the observed excess of O₂. Such calculations are characteristically performed by minimizing

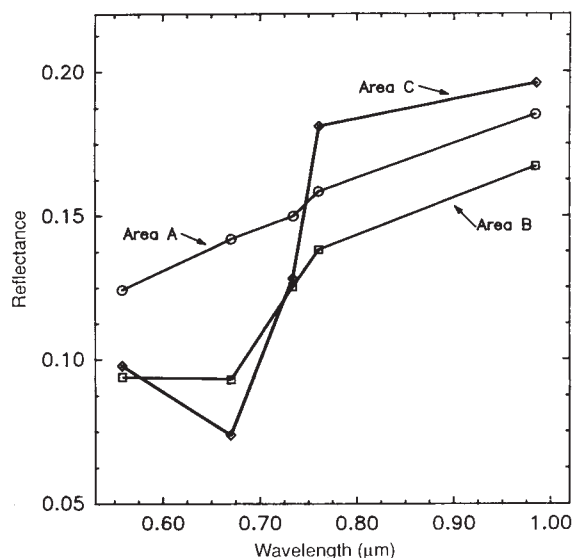


FIG. 3 Representative spectra from three areas on the land surface (see Fig. 2c). A gently sloping spectrum (circles, Area A) is consistent with any of several types of rock or soil. An intermediate spectrum (squares, Area B) shows some evidence of an absorption band near 0.67μ m (RED). Substantial areas on the surface have an unusual spectrum (diamonds, Area C) with a strong absorption in the RED band and a steep band edge just beyond 0.7μ m. This spectrum is inconsistent with all likely rock and soil types, and is plausibly associated with photo-synthetic pigments (see text).

the free energy of the system while simultaneously satisfying the equilibrium constants of all known reactions of all known reactants. This approach is less vulnerable to unknown reaction pathways or erroneous rate constants than is absolute reaction rate kinetics. But it is an idealization at best, because thermodynamic equilibrium may not be achieved in the age of the Solar System, and because there are prominent nonequilibrium processes including photochemistry and biology. Ozone is present at more than 20 orders of magnitude above its thermodynamic expectation value, but, as noted above, this disequilibrium is not considered evidence for life because a UV photochemical pathway from O₂ to this O₃ abundance is well known.

The circumstance of methane is different. It is oxidized quickly to CO₂ and H₂O, and at thermodynamic equilibrium there should not be a single methane molecule in the Earth's atmosphere^{16,17}. The disparity between observation and thermodynamic equilibrium is about 140 orders of magnitude. Clearly there is some mechanism that pumps CH₄ into the Earth's atmosphere so rapidly that substantial steady-state abundances accrue before oxidation can keep pace. It has long been suggested that an extreme disequilibrium abundance of a reduced gas such as CH₄ in an O₂-rich atmosphere could be evidence for life on Earth^{6-8,16}. The total emission flux of CH₄ into the atmosphere of the ground-truth Earth from all sources¹⁸ required to sustain the steady state abundance is $500 \pm 100 \text{ Tg yr}^{-1}$ (where 1 Tg is 10^{12} g), or $\sim 10^{-4} \text{ g cm}^{-2} \text{ yr}^{-1}$, an oxidation of all atmospheric CH₄ roughly every decade. Conceivably, methane could be injected into the Earth's atmosphere by other means (for example, the decomposition of prebiological organic matter) but so large a rate of injection, corresponding to ~ 1 bar of CH₄ after only 10^7 yr , seems highly implausible. Oxidation of such organic matter has certainly gone to completion on, for example, Venus. In fact, on the ground-truth Earth the contribution of volcanos, fumaroles, and earthquakes to atmospheric CH₄ is negligible¹⁸. About half the annual CH₄ injection is thought to arise from natural systems (methane bacteria and so on) and approximately half is anthropogenic; rice cultivation, biomass burning and flatulence from domesticated ruminants are among the principal sources¹⁸. All these sources are biological, including those deriving from fossil fuels. Thus the disequilibrium abundance of CH₄ in the Earth's oxidizing atmosphere is found by ground-based observations to be indeed caused by biology, as naive consideration of the Galileo data would suggest. It is also evidence that organic chemistry plays some role in life on Earth.

The presence of nitrous oxide (N₂O) in the atmosphere at high disequilibrium abundances is another indicator of biological processes. Nitrous oxide is partly lost to photodissociation, with an atmospheric lifetime¹⁹ of about 50 years. Although there are non-biological mechanisms for producing N₂O (for example lightning), the major source of N₂O on the ground-truth Earth is nitrogen-fixing bacteria and algae that convert soil and oceanic nitrate (NO₃⁻) to N₂ and N₂O.

Imaging

A typical Galileo image of the Earth shows continents, oceans, the Antarctic polar cap and a highly time-variable configuration of clouds. The six bands used in SSI global images can be combined in various ways to visualize specific spectral contrasts in the clouds or on the surface. Of the many possibilities, three categories of band combinations prove most informative. First is (RED, GRN, VIO) (equivalent wavelengths 0.670 , 0.558 and 0.407μ m respectively), which gives an approximately natural colour view (Fig. 2a). Large expanses of blue ocean and apparent coastlines are present, and close examination of the images shows a region of specular reflection in ocean but not on land. Clouds cover much of the land surface, but in clear areas extreme albedo contrasts are seen. The lighter areas have a colour compatible with mineral soils, while a greenish tint can be perceived in the darkest areas.

Second, we may translate (1.0, RED, GRN) (equivalent wavelength of 1.0 is 0.984 μm) bands to red, green and blue, respectively, in a false-colour image (Fig. 2b). This version reveals two kinds of sharp spectral contrasts. First, the cyan colour of the bright polar cap is caused by absorption in the 1.0 μm band, which confirms its H_2O -ice composition, and second, some land areas appear bright red, indicating a high albedo at 1.0 μm along with a low albedo in the visible spectrum (especially RED). Even without additional spectral information, a GRN/RED/1.0 signature so extreme is inconsistent with most dry or hydrated rocks or soils that might be expected on the surface of a terrestrial planet.

Using the 0.73 and 0.76 μm bands deepens the mystery. Figure 3 shows average mean spectra for three classes of land surface. (VIO is not plotted because of strong, differential atmospheric scattering corrections that are not fully modelled). The spectrum of area A (circles) has a gentle slope that rises in brightness uniformly from GRN to 1.0 μm , consistent with a variety of dark rock or mineral-soil surfaces. The spectrum of area B, an intermediate-albedo area, shows an absorption centred at RED. Although such areas are common on the surface, no common igneous, altered igneous, or sedimentary rock/soil surface displays such a signature²⁰. The spectrum of area C is from a large low-albedo area in the northern part of the continent. Its very unusual spectrum has a strong absorption in the visible spectrum and a very steep spectral slope from 0.67 (RED) to 0.76 μm .

The third category of colour composite uses three closely spaced bands (0.76, 0.73, RED), and emphasizes strong spectral slopes over this narrow wavelength range (0.67–0.76 μm) (Fig. 2c). Areas that have gently sloped spectra consistent with plausible rocks and soils appear grey here, whereas areas dominated by the unusual surface material with a strong absorption edge near 0.7 μm appear in greenish-yellow/orange hues. (The only other substantial contrast are the magenta areas in clouds, due to the 0.73 μm water-vapour band.)

These spectra cannot be uniquely interpreted; however, because they are inconsistent with any known rock or soil types on terrestrial planets of iron silicate surface composition, with or without aqueous alteration,²⁰ unusual materials must be considered. The possibility naturally arises that the strong RED

absorption is the signature of a light-harvesting pigment in a photosynthetic system—already suggested, as discussed above, as one possible explanation of the large atmospheric O_2 abundance. Plant life might have to be widespread in order to sustain this O_2 abundance in the face of likely loss by oxidation of the crust. Substantial areas of the surface seem to be covered by this pigment, which on the ground-truth planet is of course chlorophyll *a* and *b*. Photosynthesis might also be driven by oceanic microbes and plants, but the strong red-to-infrared absorption of water makes their pigments much more difficult to detect.

Morphology and topographic resolution

During this fly-by, Galileo's highest resolution systematic imaging of Earth was obtained at 1–2 km per pixel. A two-part mosaic of the ~2,900 by 4,000 km continent of Australia was produced: the eastern half of the continent was imaged mostly with four-band (VIO, GRN, RED, 0.73) coverage at ~1 km per pixel, and the western half with three-band (VIO, GRN, RED) coverage. A mosaic of Antarctica was also produced in three bands (VIO, GRN, RED) at ~2 km per pixel (at the pole).

A later six-band global imaging sequence at ~26 km per pixel showed Australia to be dominated by large central and western deserts, but with some 1.0 μm -bright areas (presumably plant life) concentrated toward the eastern and northern coasts, while Antarctica was found to be almost entirely covered by H_2O ice (Fig. 2b). In ~1 km per pixel Galileo mosaics of the Australian continent, no reworking of the surface into geometric patterns, or other compelling indications of artefacts of a technical civilization could be discerned. Although we find some large-scale albedo boundaries in Australian coastal regions that can be associated *a posteriori* with agriculture, in our judgement they are not sufficiently distinctive to be, by themselves, indicative of intelligent life. Kilston *et al.*²¹ have demonstrated from a large array of daytime orbital imagery that very few images of the ground-truth Earth at resolution ~1 km reveal evidence of life.

In a study using daytime satellite imaging at 0.1 km resolution, C.S. and Wallace²² concluded that the chance of finding convincing artifacts in a random frame was only $\sim 10^{-2}$. They further estimated, as a function of resolution, the threshold (in terms of fraction of the ground-truth surface image) for detection of

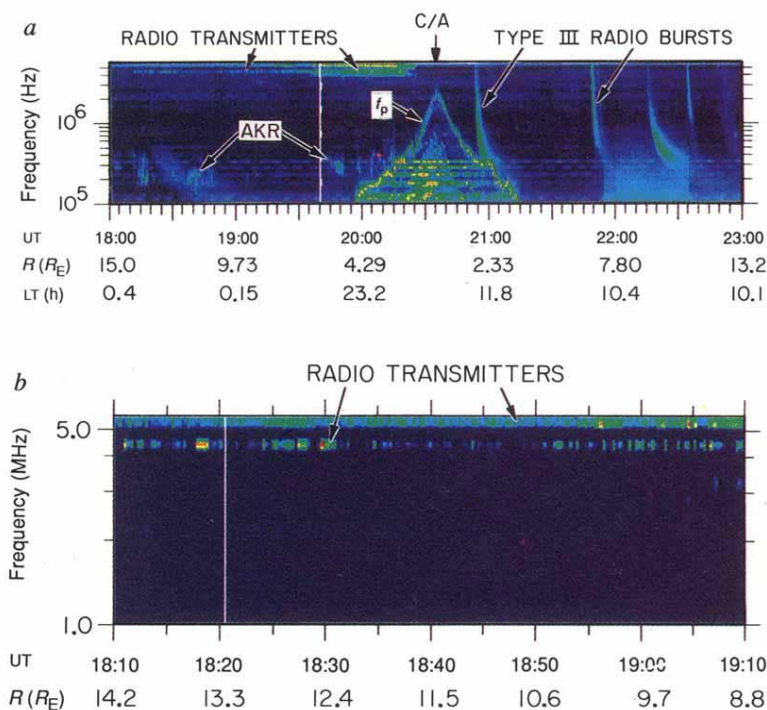


FIG. 4 A frequency-time spectrogram of the radio signals detected by the Galileo plasma wave instrument. The intensities are coded in the sequence blue-green-yellow-red, with blue lowest and red highest. Several natural sources of radio emission are shown in *a*, including auroral kilometric radiation (AKR). Modulated emission at $f > 4$ MHz is shown with an expanded time scale in *b*. Modulated patterns of this type are characteristic of the transmission of information, and would be highly unusual for a naturally occurring radio source. (UT, universal time; R is distance of Galileo from Earth in units of Earth's radius, R_E ; LT, local time.)

a civilization at the current human level of development (type A) and a hypothetical civilization that reworks its planetary surface to a significantly greater degree (type B). Their Fig. 29 predicts that even with complete surface coverage, detection of a type A civilization would require ~ 2 -km resolution, and type B, ~ 10 -km resolution. Galileo's Australia mosaic represents 2.3% of the whole surface of the Earth imaged at ~ 1 km per pixel, and the Antarctic mosaic adds another 4% at ~ 2 km per pixel. Unluckily, these are among the most sparsely settled regions of the ground-truth planet. Imaging about 1% of the surface at this resolution would detect a type B civilization, but imaging nearly all the surface would be required to detect one of type A²². A type B civilization is either wholly absent, or present in other continents than Australia and Antarctica. Extensive imaging of the Earth at resolution better than 0.1 km would have readily detected signs of life²³. The foregoing analysis is based on visible and near-infrared Galileo imaging in reflected sunlight. No usable imaging data were obtained from the night hemisphere of the planet.

Radio emission

Ground-truth television and radar transmissions are predicted to generate a non-thermal radio emission spectrum so strong and striking as to announce the presence of intelligent life not just over interplanetary, but over interstellar distances^{24,25}. With a rough knowledge of the composition and structure of the Earth's atmosphere and the solar UV spectrum, it is a straightforward matter²⁶ to derive the order of magnitude of N_e , the ionospheric electron density in the E and F layers, yielding a plasma frequency $f_p \sim 1$ –5 MHz. The higher value is the maximum for the sunlit hemisphere. Radio emissions from a technical civilization at the Earth's surface should be detectable only at $f > f_p$.

During the Galileo fly-by, the plasma wave instrument detected radio signals, plausibly escaping through the nightside ionosphere from ground-based radio transmitters. Of all Galileo science measurements, these signals provide the only indication of intelligent, technological life on Earth. They are illustrated in the PWS frequency–time spectrograms in Fig. 4. The transmitter signals can be seen near the top of the spectrograms in Fig. 4a, starting at about 18:00 UT and extending to about 20:25 UT, shortly before closest approach (C/A). These signals were detected only during the inbound, nightside pass, and not on the outbound, dayside, pass. Also seen in Fig. 4a are a series of Type III solar radio bursts²⁷ and several bursts of auroral kilometric radiation²⁸. The narrow-band emission line labelled f_p is an electrostatic oscillation excited at the local electron plasma frequency by thermal fluctuations in the ionospheric plasma. The sharp peak in f_p near closest approach is caused by the local peak in N_e as the spacecraft passed through the ionosphere at an altitude of ~ 950 km.

The identification of the narrow-band emissions near 4–5 MHz as ground-based radio transmitters is based on studies by LaBell *et al.*²⁹ and Keller³⁰ of the ground-truth planet; satellite observations clearly showed the signals originating from surface transmitters. That a similar conclusion could be drawn from the aforementioned Galileo fly-by data alone, without any previous knowledge of the existence of radio transmitters on Earth, is plausible but not beyond doubt. The rapid increase in the signal strength as the spacecraft approaches the Earth clearly indicates a near-Earth origin. The fact that the signals are only observed on the nightside inbound pass, where the ionosphere propagation cutoff frequency is sufficiently low to allow the signals to escape through the ionosphere, but not on the dayside where the cutoff prevents escape, strongly indicates a source beneath the ionosphere. Taking into account the expected exponential decrease in ionospheric N_e with increasing altitude, the theoretical maximum $f_p \sim 5$ MHz is in good agreement with the peak local plasma frequency (~ 2 MHz) observed at C/A, which was over the sunlit hemisphere at 16:30 LT. (These data might also

be consistent with a class of nocturnal transmitting stations orbiting the Earth above the ionosphere, but the conclusion of intelligent life would be unchanged.)

More difficult to estimate from first principles is the nightside f_p , because it depends on various losses and secondary ionization processes that cannot easily be evaluated from Galileo data alone. But because the primary (UV) ionization source is absent on the nightside, it is clear that N_e there will be much lower than on the dayside. From remote observations of Venus and Mars, it is entirely plausible that the nightside f_p should be depressed by factors of three to ten, relative to dayside values, permitting the escape of radio signals through the nightside ionosphere at 4–5 MHz, and providing a consistent overall interpretation of the day–night asymmetry in the observed radio signal. Ground-truth measurements of the night-time ionosphere show that $f_p < 4$ MHz is quite common, particularly at high latitudes³¹.

The signals are confined mainly to two or three distinct channels near the top of the spectrogram (Fig. 4). The fact that the central frequencies of these signals remain constant over periods of hours strongly suggests an artificial origin. Naturally generated radio emissions almost always display significant long-term frequency drifts. Even more definitive is the existence of pulse-like amplitude modulation. When the spectrum in Fig. 4a is expanded (Fig. 4b), the individual narrow-band components can be seen to have a complex modulation pattern. Although the time resolution of the instrument (18.67 s) is inadequate to decode the modulation, such modulation patterns are never observed for naturally occurring radio emissions and implies the transmission of information. On the basis of these observations, a strong case can be made that the signals are generated by an intelligent form of life on Earth.

Conclusions

From the Galileo fly-by, an observer otherwise unfamiliar with the Earth would be able to draw the following conclusions: The planet is covered with large amounts of water present as vapour, as snow and ice, and as oceans. If any biota exists, it is plausibly water-based. There is so much O_2 in the atmosphere as to cast doubt on the proposition that UV photodissociation of water vapour and the escape of hydrogen provide an adequate source. An alternative explanation is biologically mediated photodissociation of water by visible light as the first step in photosynthesis. An unusual red-absorbing pigment that may serve this purpose, corresponding to no plausible mineral, is found widely on land. Methane is detected at ~ 1 p.p.m., some 140 orders of magnitude higher than the thermodynamic equilibrium value in this oxygen-rich atmosphere. Only biological processes are likely to generate so large a disparity. But how plausible a world covered with carbon-fixing photosynthetic organism, using H_2O as the electron donor and generating a massive (and poisonous) O_2 atmosphere, might be to observers from a very different world is an open question.

Several per cent of the land surface was imaged at a resolution of a few kilometres—entirely in Australia and Antarctica. No unambiguous sign of technological geometrization was found. Narrow-band pulsed amplitude-modulated radio transmission above the plasma frequency strongly suggests the presence of a technological civilization. Most of the evidence uncovered by Galileo would have been discovered by a similar fly-by spacecraft as long ago as about 2 billion (10^9) years. In contrast, modulated narrow-band radio transmissions could not have been detected before this century.

The identification of molecules profoundly out of thermodynamic equilibrium, unexplained by any non-biological process; widespread pigments that cannot be understood by geochemical processes; and modulated radio emission are together evidence of life on Earth without any *a priori* assumptions about its chemistry; $200 \text{ g cm}^{-2} O_2$ is at least suggestive of biology. Negative results in spacecraft exploration of other planets have thus much wider application than merely to

'life as we know it'. Of course putative ecosystems that are only weakly coupled to the surface environment (for example, subsurface^{32,33}) are not excluded by such observations.

Surface oceans of liquid water are unique in the Solar System; conceivably this is part of the reason that Earth is the only planet in the Solar System with abundant surface life. Similar spectroscopic methods, although without resolution of the disk, may be useful in examining planets of other stars for indigenous life, as has been suggested for O₂ by Owen.³⁴ Large filled-aperture or interferometric telescopes intended for investigating extrasolar planets, especially those of terrestrial mass, might incorporate into their design the wavelength range and spectral resolution that has proved useful in the present work. In the radio search for extraterrestrial intelligence (SETI), optimum

transmission through interstellar space and plausible planetary atmospheres lies in the 1–3 GHz range, not the MHz range used here.

The Galileo mission constitutes an apparently unique control experiment on the ability of fly-by spacecraft to detect life at various stages of evolutionary development on other worlds in the Solar System. Although a similar opportunity arises in the summer 1999 Earth fly-by of the ESA/NASA Cassini spacecraft on its way to Saturn, there are, because of funding constraints, no plans to observe the Earth with Cassini. Although a great deal more exploration remains to be done before such conclusions can be considered secure, our results are consistent with the hypothesis that widespread biological activity now exists, of all the worlds in this Solar System, only on Earth. □

Received 17 February; accepted 14 September 1993.

1. Carlson, R. W. *et al. Space Sci. Rev.* **60**, 457–502 (1992).
2. Hord, C. W. *et al. Space Sci. Rev.* **60**, 503–530 (1992).
3. Belton, M. J. S. *et al. Space Sci. Rev.* **60**, 413–455 (1992).
4. Gurnett, D. A. *et al. Space Sci. Rev.* **60**, 341–355 (1992).
5. Lederberg, J. *Nature* **207**, 9–13 (1965).
6. Lovelock, J. *Nature* **207**, 568–570 (1965).
7. Lovelock, J. E. *Proc. R. Soc. B* **189**, 167–181 (1975).
8. Sagan, C. *Proc. R. Soc. B* **189**, 143–166 (1975).
9. Carlson, R. W., Arakelian, T. & Smythe, W. D. *Antarct. J. U.S.* (in the press).
10. Drossart, P. J. *et al. Planet. Space Sci.* (submitted).
11. Henderson, L. J. *The Fitness of the Environment: An Inquiry into the Biological Significance of the Properties of Matter* (Peter Smith, Gloucester, Mass., 1913).
12. von Zahn, U., Kumar, R. S., Neimann, H. & Prinn, R. in *Venus Ch. 13* (eds Hunten, D. M., Colin, L., Donahue, T. M. & Moroz, V. I.) (Univ. of Arizona Press, Tucson, 1983).
13. Owen, T. in *Mars Ch. 25* (eds Keiffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) (Univ. of Arizona Press, Tucson, 1992).
14. Walker, J. G. C. *Evolution of the Atmosphere* (Macmillan, New York, 1977).
15. Léger, A., Pirre, M. & Marceau, F. *J. Astr. Astrophys.* (in the press).
16. Lippincott, E. R., Eck, R. V., Dayhoff, M. O. & Sagan, C. *Astrophys. J.* **147**, 753–764 (1967).
17. Chameides, W. L. & Davis, D. D. *Chem. Engng. News* **60**, 38–52 (1992).
18. Hogan, K. B., Hoffman, J. S. & Thompson, A. M. *Nature* **354**, 181–182 (1991).
19. Lewis, J. S. & Prinn, R. G. *Planets and Their Atmospheres* (Academic, New York, 1984).
20. Bowker, D. E., Davis, R. E., Myrick, D. L., Stacy, K. & Jones, W. T. *Ref. Publ. No. 1139* (NASA, Washington, 1985).

21. Kilston, S. D., Drummond, R. R. & Sagan, C. *Icarus* **5**, 79–98 (1966).
22. Sagan, C. & Wallace, D. *Icarus* **15**, 515–554 (1971).
23. Sagan, C. *et al. in Biology and the Exploration of Mars Ch. 9* (eds Pittendrigh, C. S., Vishniac, W. & Pearman, J. P. T.) (Natn. Acad. Sci. Washington, 1966).
24. Shklovskii, I. S. & Sagan, C. *Intelligent Life in the Universe* (Holden Day, San Francisco, 1966).
25. Sullivan, W. T., Brown, S. & Wetherill, C. *Science* **199**, 377–388 (1978).
26. Chapman, S. *Proc. phys. Soc.* **43**, 483–501 (1931).
27. Fainberg, J. & Stone, R. G. *Space Sci. Rev.* **16**, 145–188 (1974).
28. Gurnett, D. A. *J. geophys. Res.* **79**, 4227–4238 (1974).
29. LaBelle, J., Trumann, R. A., Boehm, M. H. & Gewecke, K. *Radio Sci.* **24**, 725–737 (1989).
30. Keller, A. thesis, Univ. Iowa (1990).
31. Hines, C. O., Paghis, I., Hartz, T. R. & Fejer, J. A. *Physics of the Earth's Upper Atmosphere* (Prentice Hall, Englewood Cliffs, 1965).
32. Lederberg, J. & Sagan, C. *Proc. Natn. Acad. Sci. U.S.A.* **48**, 1473–1475 (1962).
33. Gold, T. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6045–6049 (1992).
34. Owen, T. in *Strategies for the Search for Life in the Universe* (ed. Papagiannis, M.) 177–185 (Reidel, Dordrecht, 1980).

ACKNOWLEDGEMENTS. We are grateful to W. O'Neil, F. Fanale, T. Johnson, C. Chapman, M. Belton and other Galileo colleagues, as well as W. Sullivan, for encouragement and support; and to J. Lederberg, A. Léger, J. Lovelock, A. McEwen, T. Owen and J. Tarter for comments. This research was supported by the Galileo Project Office, Jet Propulsion Laboratory, NASA and by a grant from NASA's Exobiology Program.

Direct observation of kinesin stepping by optical trapping interferometry

Karel Svoboda^{*†}, Christoph F. Schmidt^{*‡}, Bruce J. Schnapp[§]
& Steven M. Block^{*||}

^{*} Rowland Institute for Science, 100 Edwin Land Boulevard, Cambridge, Massachusetts 02142, USA

[†] Committee on Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA

[§] Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA

Do biological motors move with regular steps? To address this question, we constructed instrumentation with the spatial and temporal sensitivity to resolve movement on a molecular scale. We deposited silica beads carrying single molecules of the motor protein kinesin on microtubules using optical tweezers and analysed their motion under controlled loads by interferometry. We find that kinesin moves with 8-nm steps.

ENZYMES such as myosin, kinesin, dynein and their relatives are linear motors converting the energy of ATP hydrolysis into mechanical work, moving along polymer substrates: myosin along actin filaments in muscle and other cells; kinesin and dynein along microtubules. Motion derives from a mechanochemical cycle during which the motor protein binds to successive sites along the substrate, in such a way as to move forward on average^{1,3}. Whether this cycle is accomplished through a

swinging crossbridge, and how cycles of advancement are coupled to ATP hydrolysis, have been the subject of considerable debate^{4–6}. *In vitro* assays for motility^{7,8}, using purified components interacting in well-defined experimental geometries, permit, in principle, measurement of speeds, forces, displacements, cycle timing and other physical properties of individual molecules, using native or mutant proteins^{9–14}.

Are there steps?

Do motor proteins make characteristic steps? That is, do they move forward in a discontinuous fashion, dwelling for times at

[‡] Present address: Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA.

^{||} To whom correspondence should be addressed.

well-defined positions on the substrate, interspersed with periods of advancement? We define the step size, which may be invariant or represent a distribution of values, as the distance moved forwards between dwell states. (Here we use 'step size' in its physical sense, and not to mean the average distance moved per molecule of ATP hydrolysed, also termed the 'sliding distance'^{4,9,11,15,18}. The latter corresponds to the physical step size, or an integral multiple thereof, in models in which ATP hydrolysis and stepping are tightly coupled.) If stepping occurs, the distributions of step sizes and dwell times will place constraints on possible mechanisms for movement.

The motor protein kinesin has advantages over myosin for physical studies, despite a comparative dearth of biochemical data. It can be made to move slowly, permitting better time-averaging of position, and it remains attached to the substrate for a substantial fraction of the kinetic cycle^{12,13,19}, reducing the magnitude of brownian excursions. Movement by single molecules of kinesin has been demonstrated^{12,13}, and kinesin can transport small beads, which provide high-contrast markers for motor position in the microscope⁸. Unlike actin filaments, microtubules are relatively rigid²⁰, and can be visualized by video-enhanced differential interference-contrast microscopy²¹ (DIC). Finally, recombinant kinesin expressed in bacteria has been shown to move *in vitro*, paving the way for future study¹⁴.

Significant technical difficulties nevertheless exist in measuring movements of single molecules. The motions occur on length scales of ångströms to nanometres and on timescales of milliseconds and less. To obtain the high spatial and temporal sensitivity required, we combined optical tweezers^{22,23} with a dual-beam interferometer²⁴ to produce an 'optical trapping interferometer'. A laser, focused through a microscope objective of high numerical aperture, provides position detection and trapping functions simultaneously, and can produce controlled, calibratable forces in the piconewton range. Using this device, we captured silica beads with kinesin molecules bound to their surface out of a suspension and deposited them onto microtubules immobilized on a coverslip. We then observed the fine structure of the motion as beads developed load by moving away from the centre of the trap.

Our data provide evidence for steps under three sets of conditions. At moderate levels of ATP and low loads, kinesin movement was load-independent, and the relatively high speed, in combination with brownian motion, precluded direct visualization of steps. A statistical analysis of the trajectories, however, revealed a stepwise character to the motion. At saturating levels of ATP and high loads, or at low levels of ATP and low loads, (when movement is slowed mechanically or chemically), it was possible to see the abrupt transitions directly. We estimate the step size to be 8 nm, a distance that corresponds closely to the spacing between adjacent α - β tubulin dimers in the protofilament of the microtubule²⁵, suggesting that the elementary step spans one dimer.

Optical trapping interferometer

Polarized laser light is introduced at a point just below the objective Wollaston prism into a microscope equipped with DIC optics (Fig. 1a). The prism splits the light into two beams with orthogonal polarization: these are focused to overlapping, diffraction-limited spots at the specimen plane, and together they function as a single optical trap²³. A phase object in the specimen plane inside the region illuminated by the two spots (the detector zone) introduces a relative retardation between the beams, so that when they recombine and interfere in the condenser Wollaston prism, elliptically polarized light results (Fig. 1a, left). The degree of ellipticity is measured by additional optics²⁴ and provides a sensitive measure of retardation, which changes during movement, passing through zero when the object exactly straddles the two spots (for example, a bead moved along a microtubule, as shown in Fig. 1a, right). For small excursions

(out to ~ 150 nm), the output of the detector system is linear with displacement (Fig. 1a, inset).

Over most of its bandwidth, detector noise is at or below $1 \text{ Å}/\sqrt{\text{Hz}}$ (Fig. 1b). The response to a 100 Hz sinusoidal calibration signal of 1 nm amplitude (Fig. 1b, inset) shows this ångström-level noise. An optical trapping interferometer has advantages over conventional split-photodiode systems^{26,27}. Because it is a non-imaging device, it is relatively insensitive to vibrations of the photodetector. Laser light levels ensure that detectors do not become shot-noise-limited. The detector zone can be repositioned rapidly within the microscope field of view. Because trapping and position-sensing functions are provided by the same laser beam, the two are intrinsically aligned. Finally, the arrangement does not interfere with the simultaneous use of conventional DIC imaging.

Trapping force can be calibrated in a number of ways^{28–31}. Two independent methods were used here. A rapid, convenient method applicable for small excursions from the trap centre is to measure the thermal motion of an unbound, trapped bead (Fig. 1c). The optical trap behaves like a linear spring, so that dynamics correspond to brownian motion in a harmonic potential, which has a lorentzian power spectrum. Experimental spectra are well fitted by lorentzians, and the corner frequency provides the ratio of the trap spring constant to the frictional drag coefficient of the bead³². The latter is obtained from Stokes' law, after correction for the proximity of the coverslip surface³³. This approach permits forces on individual beads to be characterized *in situ*, just before or after their use in motility assays. A second approach, useful for larger excursions, is to move the stage in a sinusoidal motion while recording bead displacement from the fixed trap position (details are given in Fig. 1 legend). Both methods give results agreeing within 10%.

Low-load regime

Silica beads were incubated with small amounts of kinesin, such that they carried fewer than one active molecule, on average, and were deposited on microtubules with the optical trap. The power was set to ~ 17 mW at the specimen plane, providing a nominal trapping force that varied linearly from 0 pN at the centre to ~ 1.5 pN at the edge of the detector zone (~ 200 nm). The ATP concentration was fixed at $10 \mu\text{M}$. Mean bead velocity in the outermost region of the trap, corresponding to the greatest force, was within 10% of that measured near the trap centre (51 nm s^{-1} versus 54 nm s^{-1}). As observed previously¹³, beads frequently released from the microtubule after a variable period of progress (runs) and were drawn rapidly back to the trap centre (within ≤ 2 ms), whereupon they rebound and began to move again (Fig. 2a). Multiple cycles of attachment, movement and release were observed, with the same bead passing repeatedly through the detector region (as many as 20 times). The points of release, hence the corresponding force levels, were not the same at each pass. Measurements were terminated when a bead travelled out of the trap altogether or became stuck.

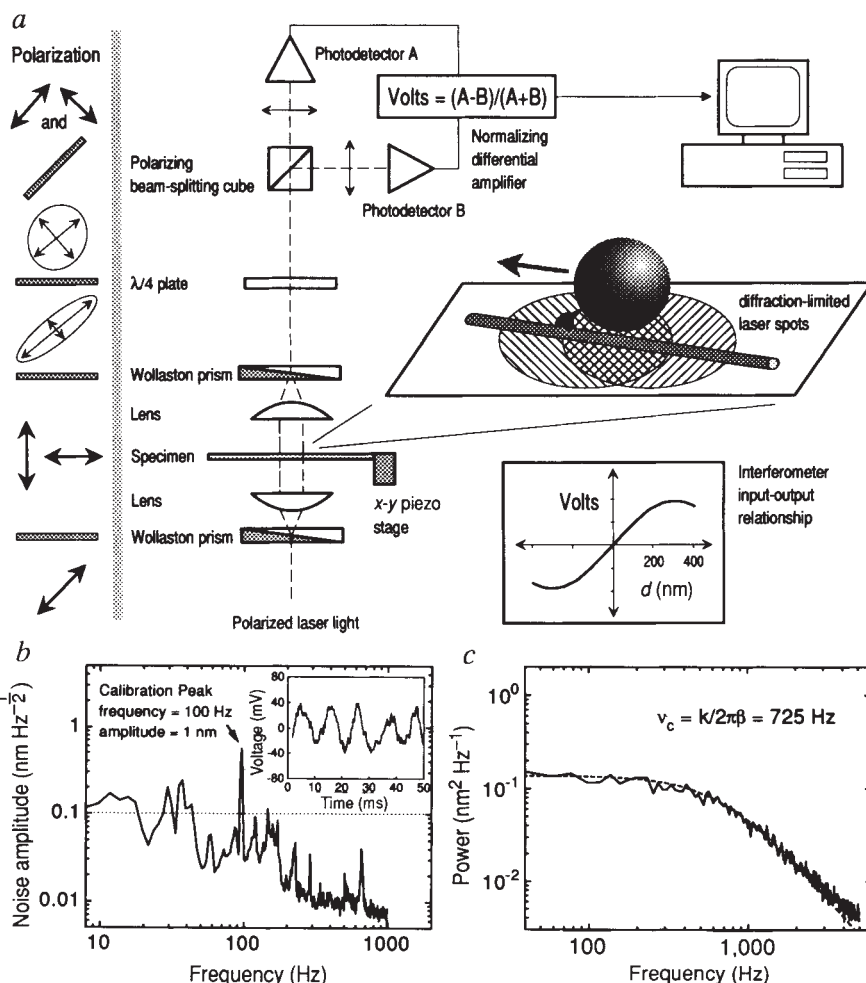
Bead displacement during one run is shown in Fig. 2b. The thermal noise in the displacement signal decreased as the bead developed load and walked toward the edge of the trap (Fig. 2c), indicating a nonlinear elasticity. This can be explained as follows. The bead's position is determined by its linkage to two springs. The first is the optical trap, with spring constant k_{trap} . The second is the linkage connecting the bead to the microtubule, acting through the motor, with spring constant k_{motor} . Both springs are extended as the motor pulls forward. If these springs were linear, thermal motion would be independent of the equilibrium position, with a mean-square displacement given by $\langle x^2 \rangle = k_B T / (k_{\text{trap}} + k_{\text{motor}})$, where k_B is Boltzmann's constant and T is the temperature. Therefore, the data of Fig. 2c imply that k_{motor} gets stiffer with increased extension. It is this diminution of noise, in part, that makes it possible to see steps: when the amplitude of the thermal noise is substantially larger than the step size, steps cannot be detected by any of the methods

used here (our unpublished computer simulations). This underscores the need to keep the bead-to-motor linkage stiff in experiments designed to look for molecular-scale motion, and also for imposing a stiff external tension. Considerable variability was observed in the mean noise level from bead to bead in these

experiments, as well as in the degree of noise reduction under tension (by factors of 1–3). For this reason, and because of the relatively large extension (tens of nanometres), we consider it unlikely that stretching of the kinesin molecule itself was chiefly responsible for nonlinear behaviour. One reasonable explanation

FIG. 1 The optical trapping interferometer. **a** (left), The diagram illustrates the polarization state of light as it passes through elements of the system, viewed along the optical axis. **a** (right), A schematic of the instrument. Polarized laser light passes through a Wollaston prism and is focused to two overlapping diffraction-limited spots ($\sim 1\ \mu\text{m}$ diameter) with orthogonal polarization, separated by roughly $\sim 250\ \text{nm}$. After passage through the specimen (a bead propelled along a microtubule by a kinesin molecule), light recombines in the upper Wollaston and develops slightly elliptical polarization (see text). Ellipticity is measured by a quarter waveplate, which produces nearly circularly polarized light, followed by a polarizing beam-splitting cube, which splits the light into two nearly equal components. The difference in intensity is detected by photodiodes and a normalizing differential amplifier. Signals were analysed offline (LabView, National Instruments). **b**, Sensitivity of the interferometer. The graph shows the spectral noise density of the interferometer responding to a 100-Hz calibration signal of 1-nm amplitude. The large peak (arrow) corresponds to the signal. The detector voltage output (inset) shows both signal and noise. **c**, Force calibration of the optical trap. The thermal noise spectrum of a bead trapped using 58 mW of laser power is shown (solid line). The spectrum is fitted by a lorentzian (dashed line).

METHODS. We used a modified inverted microscope (Axiovert 35, Carl Zeiss) equipped with Nomarski DIC optics (Plan Neofluar 100 $\times$ /1.3NA oil objective) fixed to a vibration isolation table (TMC Corp.). Light from a Nd:YLF laser (CW, 3W TEM₀₀, $\lambda = 1,047\ \text{nm}$; Spectra Physics) was coupled by an optical fibre. Beam-steering was accomplished with a telescope arrangement^{13,23}. An x-y piezo stage (Physik Instrumente) allowed positioning of the specimen under computer control. To measure instrument noise, a bead was embedded in polyacrylamide to suppress brownian motion and introduced into the detector. The stage was moved in a sinusoidal motion along the x axis to provide a calibration signal while recording voltage, and the power spectrum was computed. The power spectrum was scaled by computing the integral under the calibration peak and setting this equal to the mean-square value of the sinusoidal displacement amplitude. To calibrate the trap stiffness, and thereby the force, a bead in solution was trapped (typically $\sim 2\ \mu\text{m}$ above the coverslip) and its brownian motion recorded, from which a power spectrum was computed. The corner frequency provides the ratio of trap stiffness to the viscous drag of the bead; drag was calculated from the bead's diameter and corrected for proximity to the coverslip. One-sided power spectra were normalized such that $\langle x^2 \rangle = \int_0^\infty S(f) df$, where $\langle x^2 \rangle$ is the mean-square displacement and $S(f)$ is the power at a given frequency. Trapping force changed by less than 10% as the distance from the coverslip was reduced from 2 to 1 μm (data not shown). The force profile of the trap was also mapped over a larger range of displacements by moving the stage in sinusoidal fashion (amplitude $A = 2.5\ \mu\text{m}$ at frequency f) while monitoring the peak displacement, y , of a bead. At low frequencies, the force exerted by the fluid is $F = 2\pi\eta\beta f = k_{\text{trap}}y$, where $\beta = 5.7 \times 10^{-6}\ \text{pN s nm}^{-1}$ is the drag coefficient of the bead. The proportionality between y and f gave the mean stiffness, $(4.3 \pm 0.3) \times 10^{-4}\ \text{pN nm}^{-1}\ \text{mW}^{-1}$, which was approximately constant out to displacements of $\pm 200\ \text{nm}$ (data not shown). The foregoing stiffness was used to compute mean forces in the outermost regions of the trap. Power at the specimen plane was estimated by measuring the external power with a meter and applying an attenuation factor of



58%, corresponding to the transmittance of the objective at 1,064 nm, determined separately (data not shown). Kinesin was purified from squid optic lobe by microtubule affinity⁴⁷ and tubulin from bovine brain⁴⁸. Experiments were done at room temperature. Silica beads (0.6 μm diameter, $6 \times 10^{-6}\ \text{w/v}$ final concentration; Bangs Labs) were incubated for at least 5 min in buffer (80 mM PIPES, 1 mM MgCl_2 , 1 mM EGTA, pH 6.9, 50 mM KCl, 0.5 mM dithiothreitol, $50\ \mu\text{g ml}^{-1}$ filtered casein, 1–500 μM ATP, 0.5 $\mu\text{g ml}^{-1}$ phosphocreatine kinase, 2 mM phosphocreatine; in some experiments, the last three reagents were replaced by 2 mM AMP-PNP). Beads were incubated for $>1\ \text{h}$ with kinesin diluted $\sim 1:10,000$ from stock ($\sim 50\ \mu\text{g ml}^{-1}$ kinesin heavy chain). Taxol-stabilized microtubules were introduced into a flow chamber in which the coverslip had been treated with 4-aminobutyl-dimethylmethoxysilane (Huls America); microtubules bound tightly to this surface. The chamber was then incubated with 1 mg ml^{-1} casein (in 80 mM PIPES, 1 mM MgCl_2 , 1 mM EGTA, pH 6.9) for 10 min, rinsed with buffer, and kinesin-coated beads introduced. Beads were captured from solution with the trap and deposited on a microtubule selected with its long axis parallel to the Wollaston shear direction, the direction of detector sensitivity. Fewer than half the beads bound and moved when placed on a microtubule. Under these conditions, beads carry Poisson-distributed numbers of functional motors¹³. The bead size in these studies was such that the chance that any bead carried two motors in sufficient proximity to interact simultaneously with the microtubule was less than 2%, assuming, generously, that kinesin heads can reach 100 nm from their points of attachment.

for the nonlinearity is that the bead-motor linkage behaves like an entropic spring, with the swivelling motion of the bead at the end of its tether contributing degrees of freedom. This linkage becomes taut under load, such that subsequent displacements of the motor are communicated sharply to the bead.

Beads under the conditions shown in Fig. 2*a, b* moved close to 50 nm s^{-1} . The speed and the thermal noise in the data made a statistical analysis necessary to detect periodic structure. We used the following approach. First, reduced-noise segments of all runs in an experiment were selected, corresponding to movements under load, and filtered to further reduce noise (see below). Then, a histogram of all pairwise differences in displacement was computed for each record³⁴. This 'pairwise distance distribution function' (PDF) shows spatial periodicities in stepping motion, independent of times at which steps occur. Next, PDFs were averaged. Finally, the power spectrum of the average PDF was computed: this produces a peak at the mean spatial frequency of the stepper.

For a noiseless, stochastic stepper (equidistant advances after exponentially distributed time intervals), the PDF gives a set of evenly spaced peaks at multiples of the step distance. For noisy steppers, peaks become broadened, disappearing altogether when their peak widths (r.m.s. amplitudes of brownian motion) exceed the step size. The detection of steps in the presence of noise can be improved by low-pass filtering the records with an appropriate filter frequency so as to preserve stepwise character. For this purpose, we used a nonlinear median filter³⁵. Computer simulations showed that the peak stepping signal recovered from a background of gaussian noise by an appropriately chosen median filter was approximately twice the amplitude of an equivalent, linear Bessel filter (our unpublished data).

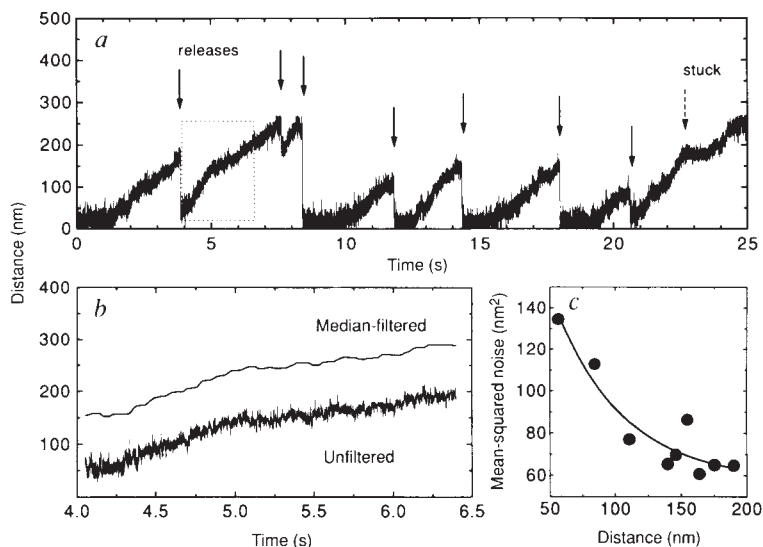
The PDF and associated power spectrum for a single run along a microtubule are shown in Fig. 3*a, b*. Even in individual runs, there was often an indication of periodicity in the PDF, with a spacing from 6–8 nm, although there was considerable

variation from run to run. Averaged data (Fig. 4*a, b*), gave peaks corresponding to an uncorrected spacing of $6.7 \pm 0.2 \text{ nm}$ (mean $\pm$ s.e.). Essentially identical data were obtained for beads moving on axonemes (data not shown). Note that backward movements, corresponding to negative distances in the PDF, were practically non-existent. Can our instrument, in combination with the analysis, reliably detect nanometre-sized steps in a noisy background? To test this, we attached beads carrying single motors to microtubules with AMP-PNP, a non-hydrolysable ATP analogue. Beads bound in rigor by AMP-PNP exhibited brownian motion but did not translocate. The piezo stage of the microscope was then stepped stochastically to move the entire specimen (bead and microtubule) through the detector zone, parallel to the long axis of the microtubule, in a series of 8-nm increments, at the same speed as the kinesin-based movement of Fig. 4*a*. The PDF and power spectrum for this experiment gave peaks at nearly the same positions (Fig. 4*c, d*). (Similarly, experiments with a stage moved in 4-nm increments gave peaks at half this spacing; data not shown.) Smaller, secondary peaks (measured relative to the sloping background) at subharmonics of the main spatial frequency are usually seen: these arise mainly from the variation in peak heights in the PDF. The measured periodicity was $6.5 \pm 0.2 \text{ nm}$, a distance that is 19% smaller than the anticipated value (8 nm). The discrepancy is a consequence of stretching in the elastic bead-motor linkage under load, causing beads to move only a fraction of the motor displacement. (The fraction is $k_{\text{motor}}/(k_{\text{trap}} + k_{\text{motor}})$ for linear springs.) This interpretation was confirmed by an experiment with beads bound directly to the coverslip that had dramatically reduced brownian motion. When the stage was again stepped stochastically by 8 nm, we obtained a strong peak at $8.0 \pm 0.2 \text{ nm}$ (Fig. 4*e, f*). Therefore, applying a 19% correction to the spectrum of Fig. 4*b* yields an adjusted estimate of $8.3 \pm 0.2 \text{ nm}$ for the mean periodicity of kinesin-based movement.

Could the filtering and/or statistical analysis produce

FIG. 2 *a*, Distance of a bead from the trap centre over 25 s. Multiple cycles of movement, release (solid arrows) and reattachment can be seen. The bead stuck briefly (<1 s) at 23 s (dashed arrow), then continued. The apparent peak between seconds 8 and 9 (third arrow) is a consequence of interferometer nonlinearity: the bead actually passed the turnover point in voltage at $\sim 280 \text{ nm}$ (Fig. 1*a*, inset) and continued its forward movement briefly before releasing. Portions of records beyond 200 nm were not analysed. *b*, Detail from the dashed box in *a*. The raw data (lower trace) were median-filtered for subsequent analysis (upper trace). *c*, The mean-square noise of the track in *b* as a function of distance from the trap centre. The noise decreases as the bead is placed under tension by the trap.

METHODS. Data were acquired at 1 kHz. A cubic polynomial was fitted to the response function (Fig. 1*a*, inset), and used to convert voltage to distance, a procedure that extended the usable instrument range to $\pm 200 \text{ nm}$ with a $\pm 5\%$ error (data not shown). The algorithm computes the absolute value of displacement, so that the distance from the centre of the trap is rectified. Record segments corresponding to movement near the trap centre were not used for analysis. The response function was determined by tracking the motion of beads stuck tightly to the coverslip surface and moved with computer-controlled voltage waveforms supplied to the piezo stage. The piezo stage was calibrated with nanometre-scale precision using video-based methods³⁸ against a 10- μm diamond-ruled grating (Donsanto Corp.). A calibration procedure was implemented that allowed beads of slightly varying size to be used, even though these scatter different amounts of light. This was made possible by the observation that the turnover point in the response function (Fig. 1*a*, inset) occurs at a fixed distance from the trap centre, independent of the bead size, and that the response function scales with voltage at that point. By measuring the turnover voltage, it was possible to establish an absolute correspondence between voltage and distance. Only runs longer than 100 nm were analysed for steps, as follows. A line was



fitted to the run, and the slope used as a first estimate of velocity, v' . A set of non-overlapping segments of duration $2d/v'$ was then fitted to the run, where d is an assumed step size, and an improved estimate of velocity, $\langle v \rangle$, was computed from these segments. The filter frequency was chosen such that $f_c \approx 2\langle v \rangle/d$. Data were filtered with a Bessel filter at $1.2f_c$ and then by a median filter³⁵ of rank r , chosen such that $f_c = 1/(2r+1)$. This procedure was determined by computer simulations to provide reliable results with stochastic steppers subject to gaussian white noise. The exact choice of d is not critical: for the final data analysis, $d=8 \text{ nm}$ was assumed, but other values (from 4 to 16 nm) gave essentially identical results. The noise in *c* is the mean-square deviation of unfiltered data from successive line segments.

artefactual results? To answer this question, we again bound beads to microtubules with AMP-PNP, but moved the specimen smoothly through the trap at the same mean speed. The PDF and associated power spectrum (Fig. 4*g, h*) had similar baselines, but no peaks corresponding to spatial periodicities. The variation in several such spectra provided a means of estimating the statistical significance of the peak in Fig. 4*b*: its likelihood of random occurrence is $P < 0.00001$.

High-load regime

The laser power was set to ~ 58 mW at the specimen plane, a level that provided a nominal force that varied linearly from 0 pN at the centre to ~ 5 pN at the edge of the detector zone. The ATP concentration was raised to 500 μ M (saturation level). When beads carrying single kinesin molecules were placed on microtubules, their motion was more erratic than at lower powers. Beads near the trap centre, experiencing low loads, moved rapidly at 300–500 nm s^{-1} . Under higher loads towards the edge of the trap, beads markedly slowed (or became stuck), although some were still able to escape altogether, opposing forces up to 5 pN. Other beads slowed or stuck before reaching the trap's edge, then continued forward at increased speed. Shuttering the trap briefly (< 0.5 s) during a slow (or stuck) episode enabled some beads, but not all, to regain forward motion. Multiple transitions between fast- and slow-moving phases were seen in individual records, with beads entering the slow phase over a range of positions (loads) in the trap. Beads still underwent

multiple cycles of movement, release and reattachment, similar to the behaviour observed at lower loads.

In the slow-moving phase, the reduced speed permitted visualization of steps. Figure 5*a, b* shows two records at high load, each containing roughly ten abrupt displacements forwards. The heights of these steps varied from ~ 5 –18 nm, with the smaller transitions averaging 8 ± 2 nm (mean $\pm$ s.d.; $n = 16$; Fig. 5*c-f*). Beads also advanced through what appeared to be 'double steps' (~ 17 nm), but it was not possible, given the bandwidth and noise in these signals, to determine whether such jumps represented two steps in rapid succession or one single step. Assuming that motors are stochastic steppers, one expects an exponential distribution of dwell times with numerous short steps. Records

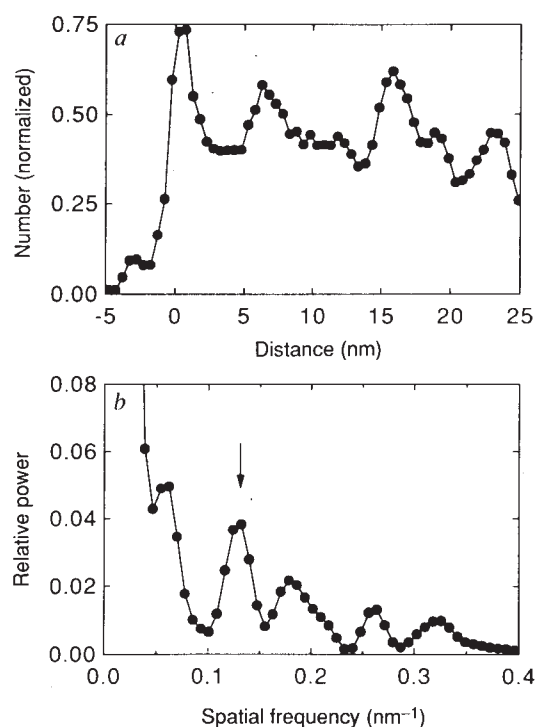


FIG. 3 *a*, The pairwise distance distribution function (PDF) for a single run. The density of interpoint distances in a median-filtered record is shown: multiple peaks with regular spacing can be identified. *b*, The normalized power spectrum of the data in *a*, showing a prominent peak at $\sim 0.130 \text{ nm}^{-1}$ (arrow), corresponding to periodicity of 7.7 nm.

METHODS. PDFs were computed by binning distance differences ($x_j - x_i$) for all ($j > i$) in a histogram, with bin width 0.5 nm. The PDF was normalized to unity at the first bin and smoothed with a 3-point moving window. The one-sided power spectrum, $S(k)$, was computed from the PDF by FFT and normalized to unity at $S(0)$. The PDF is identical to the autocorrelation function of the density of distances from the trap centre. $S(k)$ is analogous to the square of the structure factor used in X-ray scattering, with the density of distances being analogous to electron density⁴⁹.

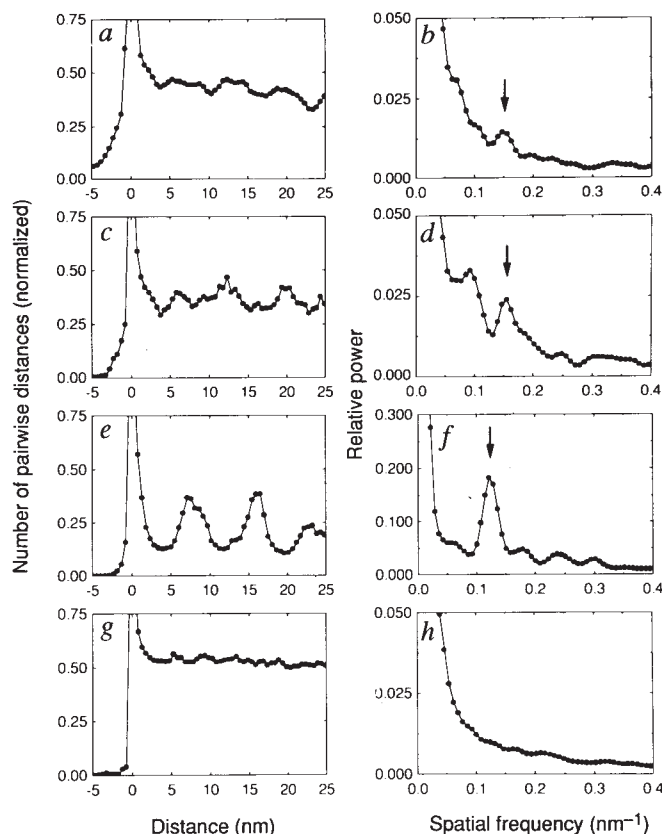


FIG. 4 *a*, The averaged pairwise distance distribution function (PDF) for kinesin movement, taken from 17 records of 10 different beads. *b*, The power spectrum of the data in *a*, with a peak at $0.149 \pm 0.004 \text{ nm}^{-1}$ (arrow), corresponding to a periodicity of 6.7 ± 0.2 nm. *c*, The averaged PDF for beads attached to the microtubule using AMP-PNP, with an 8-nm stochastically stepped stage. *d*, The power spectrum of the data in *c*, with a main peak at $0.154 \pm 0.004 \text{ nm}^{-1}$ (arrow), corresponding to a periodicity of 6.5 ± 0.2 nm. *e*, The averaged PDF for beads attached directly to the coverslip, with an 8-nm stochastically stepped stage. *f*, The spectrum of the data in *e*, with a peak at $0.125 \pm 0.003 \text{ nm}^{-1}$ (arrow), corresponding to a periodicity of 8.0 ± 0.2 nm. *g*, The averaged PDF for beads attached to the microtubule using AMP-PNP with a smoothly moved stage. *h*, The power spectrum of *g*. The average amplitude fluctuation about the mean for records of smooth movement was 0.002 units over the range of spatial frequencies 0.1–0.2; that is, there are no statistically significant peaks in this trace. All PDFs and power spectra were normalized to unity. Note different scale in *e*.

METHODS. PDFs in each panel represent averages of 17 records. Peak positions in power spectra and corresponding error estimates were determined by fitting gaussians. To estimate the statistical significance of the peak in *b*, the standard deviation per point, σ_k , for 17 spectra of individual runs of a smoothly stepped stage was computed. The likelihood of the peak was $\langle (S_k - \bar{N}_k)^2 \rangle^{1/2} / \langle \sigma_k \rangle$, where S_k is the average spectrum of kinesin-driven movement in *b*, and $\bar{N}_k$ is the average spectrum of smooth movement in *h*; averages were taken over the four spatial frequencies comprising the peak.

of the type shown in Fig. 5a, b are relatively hard to obtain: additional records must be collected before a more complete statistical analysis, including measurements of dwell-time distributions, will be possible. In parts of the record, the distance from the centre of the trap appeared to shorten briefly (~ 10 ms) just before a step (Fig. 5e, f), but this was not seen in all cases (Fig. 5c, d). We do not know if this behaviour reflects an intrinsic property of the motor, but the shortening is not due to kinesin unbinding because substrate release for periods as short as a millisecond would result in the bead being pulled off completely (Fig. 2).

Low-ATP regime

Reasoning that steps might also be resolved under low loads if beads moved slowly enough we restored the power to 17 mW and lowered the ATP concentration to 1 or 2 μM . When beads carrying single kinesin molecules were placed on a microtubule, their speeds were slow, about 5–15 nm s^{-1} , and single steps could again be seen in records that were selected for low noise (and therefore, presumably, stiffer bead-motor linkages) (Fig. 6a–e). A statistical analysis of these, like that performed in the low-load

regime, showed that steps measured 8.2 ± 1.1 nm (mean $\pm$ s.d., as determined from peak positions in power spectra for all five panels), the same size as those seen in the high-load regime. Figure 6f, g shows an example of the PDF and associated power spectrum for the data of Fig. 6e, where the mean periodicity was 8.8 ± 0.5 nm.

Is it possible that kinesin motors travel a longer, possibly variable, distance with each ATP hydrolysis, moving by some integral multiple of this spacing? That is, does a single hydrolysis result in a sequence of physical steps, rather than just one, as has been proposed for myosin^{4,5,16,36}? If this were the case for kinesin, then for extremely low concentrations of ATP (the diffusion limit), one would expect a motor to advance in a rapid burst of steps (through $\gg 8$ nm) before stopping to wait for the next ATP molecule. This would produce motion characterized by periods of zero advancement interspersed with clusters of steps. The mean rate of advance during a step series ought to be high, close to the speed for saturating levels of ATP (~ 500 nm s^{-1}). We saw no evidence for step clusters in any of our records at 1 or 2 μM ATP: overall rates of advancement were quite uniform.

Discussion

The finding of discrete stepping behaviour should allow direct measurement of kinetic parameters of the mechanochemical cycle, by determining the timing of various phases of the motion. For example, releases of the type shown in Fig. 2a permit an estimate of the cycle off-time. Once a bead near the edge of the trap detaches under low load conditions, it is returned rapidly to the centre through a distance, x , of ~ 200 nm within an average time $\tau = 1.8 \pm 0.4$ ms (mean $\pm$ s.e.). To make net forward progress, the kinesin off-time, τ_{off} , must be less than the time required for the trap to pull the bead back by one step, d , that is, $\tau_{\text{off}} < d/v = d\tau/x$, where v is the return speed at the edge of the trap. Using 8 nm for d gives $\tau_{\text{off}} \leq 72$ μs . This time is an upper limit, because velocity is practically unaffected by trapping forces. Such a short off-time signifies that kinesin somehow remains bound, sustaining load throughout most of its activity, perhaps moving hand-over-hand^{12,37}. Estimates of other kinetic parameters are ongoing subjects of investigation.

It is known that kinesin motors do not wander over the surface lattice of a microtubule, but move instead along straight paths parallel to a protofilament^{38–40}. Kinesin head fragments cloned from squid⁴¹ or *Drosophila*⁴² and expressed in bacteria have been used to decorate microtubules. Such fragments are non-functional in motility assays, but they bind microtubules and have microtubule-activated ATPase activity. Saturation binding experiments indicate a stoichiometry of one kinesin heavy-chain fragment to each tubulin dimer⁴², and crosslinking studies demonstrate that it is the β -subunit of tubulin that is bound⁴¹. Electron microscopy of decorated microtubules shows an 8-nm repeat arising from the head spacing^{41,42}. Taken together with our observations, this suggests that the two heads of a kinesin molecule walk along a single protofilament—or walk side-by-side on two adjacent protofilaments—stepping ~ 8 nm at a time, making one step per hydrolysis (or perhaps fewer, requiring multiple hydrolyses per step). If the foregoing model is correct, then during movement of single molecules, the ATP hydrolysis rate, r_{ATP} , should be related to the step size, d , and speed, v , through $r_{\text{ATP}} = v/d$. Movement at saturating levels of ATP ($v \sim 500$ nm s^{-1} and even higher²⁰) implies $r_{\text{ATP}} \geq 60$ s^{-1} mol^{-1} , 6–10-fold higher than reported^{43–45}. The hydrolysis rates for motility assays *in vitro* must therefore be substantially higher than values inferred from solution biochemistry, or motors must make several steps per ATP hydrolysed. The latter is difficult to reconcile with our data, particularly at low concentrations of ATP.

Occasionally jumps have been seen³⁸ in selected records of kinesin movement during video tracking, measuring 3.7 ± 1.7 nm, although most of the movement was subjectively

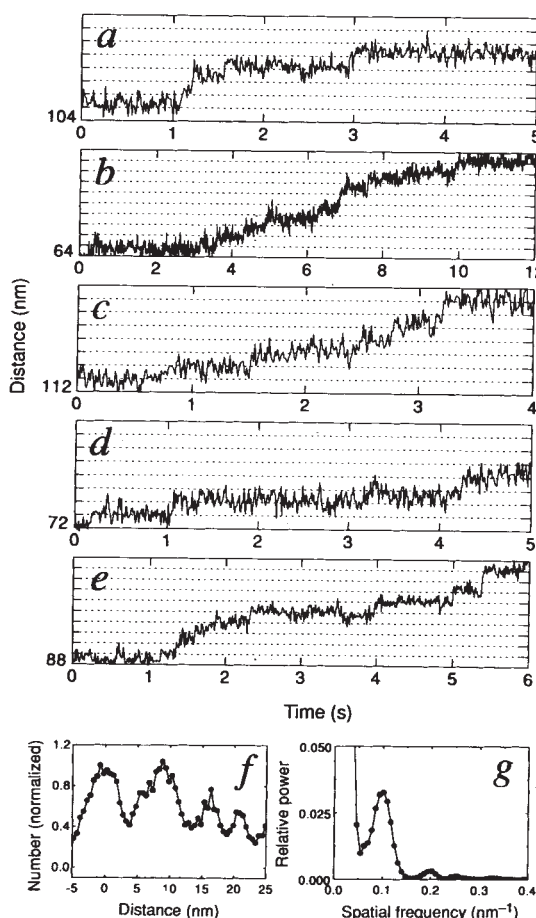
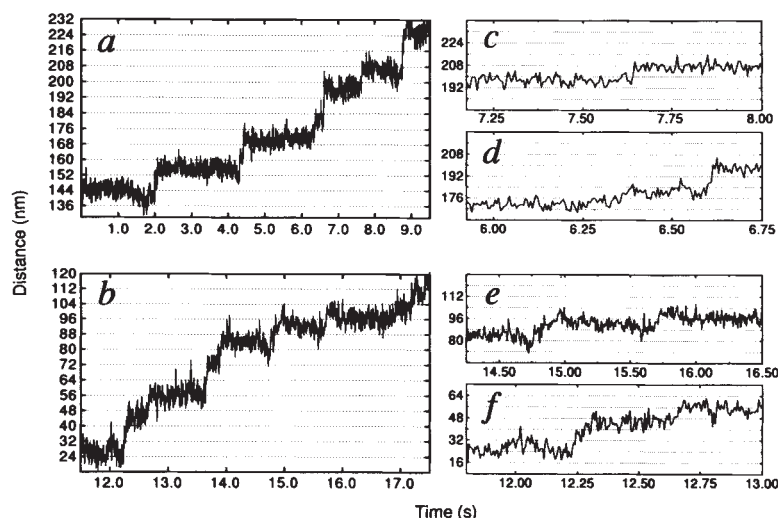


FIG. 6 a, b, Multiple steps in the displacement records of two different beads moving at an ATP concentration of 2 μM . c–e, Multiple steps in the displacement records of three different beads moving at an ATP concentration of 1 μM . Horizontal gridlines (dotted) have been drawn at a spacing of 8 nm. A 10-point jumping average of data taken at 1 kHz is plotted for an effective rate of 100 Hz. f, g, The PDF and associated power spectrum of the record in e, showing a periodicity of 8.8 ± 0.5 nm. Mean step size and s.d. for records in a–e were computed from peak positions determined from PDFs. A correction for the extension in k_{motor} has not been applied to these data, which were selected for their low noise and which presumably reflect stiffer linkages.

FIG. 5 *a, b*, Multiple steps in the displacement records of two different beads in the high-load regime. *c, d*, Details of selected steps in *a, f*, Details of selected steps in *b*. Horizontal gridlines (dotted) have been drawn at a spacing of 8 nm. A 5-point jumping average of data taken at 1 kHz is plotted for an effective rate of 200 Hz. The mean step size for these records was computed from averaged values of distance determined during segments of traces before and after transitions, identified by eye. Records at high loads were not corrected for stretching in k_{motor} because the load presumably extends that linkage to be taut.



smooth. Our instrument resolves distances of this size, but we did not find evidence for significant periodicities near ~ 4 nm. The earlier work, however, was done before the development of single-motor assays, and beads carried unknown numbers of motors, probably several. It is possible that the shorter spacing, if real, might reflect an effect of multiple heads stepping along neighbouring protofilaments.

Our data from moving beads subjected to trap-induced tensions suggest that speeds are largely unaffected by forces of 1.5 pN, and that single molecules of kinesin can still transport beads against loads up to ~ 5 pN. These findings seem incompatible with the conclusions of Kuo and Sheetz³⁰, who reported measuring the 'isometric' force generated by single kinesin molecules at 1.9 ± 0.4 pN. In their experiments, nominally arrested beads continued to move; that is, the isometric condition was not truly fulfilled, but this difficulty alone is probably insufficient to explain the discrepancy.

Finding displacement steps in the movement of kinesin molecules is in many respects analogous to detecting current steps in single-channel recordings of neurons. In both cases, steps reflect underlying motions of individual proteins, and allow one to make meaningful measurements at the level of single molecules. Before the advent of single-channel recording, Johnson (thermal) noise due to the leakage resistance of electrodes prevented the resolution of picoampere-sized current steps. Improved methods for increasing this resistance provided conditions that led directly to step detection⁴⁶. As a result, single-channel recording has produced many insights in molecular neuroscience during the past decade. For mechanoenzymes, brownian (thermal) noise of objects pulled by motors prevented the resolution of nanometre-sized displacement steps. Additional stiffness, coming in our case from the optical trap and its effect on the bead linkage, provides the required reduction in noise, permitting visualization of steps. □

Received 16 July; accepted 27 September 1993.

- Huxley, H. E. *Science* **164**, 1356–1366 (1969).
- Squire, J. *The Structural Basis of Muscle Contraction* (Plenum, London, 1981).
- Bagshaw, C. R. *Muscle Contraction* (Chapman & Hall, London, 1982).
- Burton, K. J. *Musc. Res. Cell Motil.* **13**, 590–607 (1992).
- Simmons, R. M. *Curr. Biol.* **2**, 373–375 (1992).
- Cooke, R. *CRC Crit. Rev. Biochem.* **21**, 53–117 (1986).
- Kron, S. J. & Spudich, J. A. *Proc. natn. Acad. Sci. U.S.A.* **83**, 6262–6276 (1986).
- Vale, R. D., Schnapp, B. J., Reese, T. S. & Sheetz, M. P. *Cell* **40**, 559–569 (1985).
- Ishijima, A., Doi, T., Sakurada, K. & Yanagida, T. *Nature* **352**, 301–306 (1991).
- Uyeda, T. Q. P., Kron, S. J. & Spudich, J. A. *J. molec. Biol.* **214**, 699–714 (1990).
- Uyeda, T. Q. P., Warrick, H. M., Kron, S. J. & Spudich, J. A. *Nature* **352**, 307–311 (1991).
- Howard, J., Hudspeth, A. J. & Vale, R. D. *Nature* **342**, 154–158 (1989).
- Block, S. M., Goldstein, L. S. B. & Schnapp, B. J. *Nature* **348**, 348–352 (1990).
- Yang, J. T., Saxton, W. M., Stewart, R. J., Raff, E. C. & Goldstein, L. S. B. *Science* **249**, 42–47 (1990).
- Toyoshima, Y., Kron, S. J. & Spudich, J. A. *Proc. natn. Acad. Sci. U.S.A.* **87**, 7130–7134 (1990).
- Yanagida, T., Arata, T. & Oosawa, F. *Nature* **316**, 366–369 (1985).
- Harada, Y., Sakurada, K., Aoki, T., Thomas, D. & Yanagida, T. *J. molec. Biol.* **216**, 49–68 (1990).
- Higuchi, H. & Goldman, Y. *Nature* **353**, 352–354 (1991).
- Spudich, J. A. *Nature* **348**, 284–285 (1990).
- Gittes, F., Mickey, B., Nettleton, J. & Howard, J. *J. Cell Biol.* **120**, 923–934 (1993).
- Schnapp, B. J. *Meth. Enzym.* **134**, 561–573 (1986).
- Ashkin, A., Dziedzic, J. M., Björkholm, J. E. & Chu, S. *Optics Lett.* **11**, 288–290 (1986).
- Block, S. M. in *Noninvasive Techniques in Cell Biology* 375–401 (Wiley-Liss, New York, 1990).
- Denk, W. & Webb, W. W. *Appl. Optics* **29**, 2382–2390 (1990).
- Amos, L. & Klug, A. *J. Cell Sci.* **14**, 523–549 (1974).
- Kamimura, S. *Appl. Optics* **26**, 3425–3427 (1987).
- Kamimura, S. & Kamiya, R. *J. Cell Biol.* **116**, 1443–1454 (1992).
- Block, S. M., Blair, D. F. & Berg, H. C. *Nature* **338**, 514–517 (1989).
- Ashkin, A., Schuetz, K., Dziedzic, J. M., Euteneuer, U. & Schliwa, M. *Nature* **348**, 346–348 (1990).

- Kuo, S. C. & Sheetz, M. P. *Science* **260**, 232–234 (1993).
- Simmons, R. M. et al. in *Mechanisms of Myofibril Sliding in Muscle* (eds Sugi, H. & Pollack, G.) 331–336 (Plenum, New York, 1993).
- Wang, C. W. & Uhlenbeck, G. E. in *Selected Papers on Noise and Stochastic Processes* (ed. Wax, N.) 113–132 (Dover, New York, 1954).
- Happel, J. & Brenner, H. *Low Reynolds Number Hydrodynamics* 322–331 (Kluwer Academic, Dordrecht, 1991).
- Kuo, S. C., Gelles, J., Steuer, E. & Sheetz, M. P. *J. Cell Sci. suppl.* **14**, 135–138 (1991).
- Gallagher, N. C. Jr & Wise, G. L. *IEEE Trans. Acoust. Speech Sign. Proc.* **29**, 1136–1141 (1981).
- Oosawa, F. & Hayashi, S. *Adv. Biophys.* **22**, 151–183 (1986).
- Schnapp, B. J., Crise, B., Sheetz, M. P., Reese, T. S. & Khan, S. *Proc. natn. Acad. Sci. U.S.A.* **87**, 10053–10057 (1990).
- Gelles, J., Schnapp, B. J. & Sheetz, M. P. *Nature* **331**, 450–453 (1988).
- Kamimura, S. & Mandelkow, E. *J. Cell Biol.* **108**, 865–875 (1992).
- Ray, S., Meyhoefer, E., Milligan, R. A. & Howard, J. A. *J. Cell Biol.* **121**, 1083–1093 (1993).
- Song, Y.-H. & Mandelkow, E. *Proc. natn. Acad. Sci. U.S.A.* **90**, 1671–1675 (1993).
- Harrison, B. C. et al. *Nature* **362**, 73–75 (1993).
- Kusnetsov, S. A. & Gelfand, V. I. *Proc. natn. Acad. Sci. U.S.A.* **83**, 8530–8534 (1986).
- Hackney, D. *Proc. natn. Acad. Sci. U.S.A.* **85**, 6314–6318 (1988).
- Gilbert, S. P. & Johnson, K. A. *Biochemistry* **32**, 4677–4684 (1993).
- Sakmann, B. & Neher, E. *Single-Channel Recording* (Plenum, New York, 1983).
- Weingarten, M. D., Suter, M. M., Littman, D. R. & Kirschner, M. W. *Biochemistry* **13**, 5529–5537 (1974).
- Schnapp, B. J. & Reese, T. S. *Proc. natn. Acad. Sci. U.S.A.* **86**, 1548–1552 (1989).
- Glatter, O. & Kratky, H. C. *Small Angle X-Ray Scattering* (Academic, New York, 1982).

ACKNOWLEDGEMENTS. We thank W. Denk, C. Godek, M. Meister, P. Mitra and R. Stewart for discussions and advice, W. Hill for electronic design, and R. Stewart for help with motility assays and protein purification. This work was supported by the Rowland Institute for Science (K.S., C.F.S., S.M.B.), and partial support from the NIH (K.S. and B.J.S.), the Lucille P. Markey Charitable Trust (B.J.S.), and the University of Michigan (C.F.S.).

A possible low-mass type Ia supernova

Pilar Ruiz-Lapuente*, David J. Jeffery*, Peter M. Challis*, Alexei V. Filippenko†, Robert P. Kirshner*, Luis C. Ho‡, Brian P. Schmidt*, Francisco Sánchez§ & Ramon Canal§

* Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts 02138, USA

† Department of Astronomy, and Center for Particle Astrophysics, University of California, Berkeley, California 94720, USA

‡ Instituto de Astrofísica de Canarias, La Laguna, Tenerife, Spain

§ Departament d'Astronomia i Meteorologia, Universitat de Barcelona, Barcelona, Spain

IN the standard model for type Ia supernovae¹, a massive white dwarf in a binary system accretes matter from the companion star until it reaches the Chandrasekhar mass (the stability limit for degenerate-electron stars, corresponding to ~ 1.4 solar masses), and a runaway thermonuclear explosion ensues. In a popular variant of this model², the companion star is also a white dwarf. But regardless of the nature of the companion, the invariance of the Chandrasekhar mass implies that all type Ia supernovae will be similar in luminosity³, making them ideal 'standard candles' for determining extragalactic distances, and hence the Hubble constant. In the context of the standard model, the recent type Ia supernova SN1991bg is hard to explain: it was underluminous at all observed epochs, leading to suggestions^{4,5} that the mass of the progenitor was unusually low. Here we present model calculations, based on more recent spectra, which point to a mass of the white dwarf of ~ 0.7 solar masses—well below the Chandrasekhar mass. Moreover, the late spectrum shows evidence of emission from low-velocity hydrogen gas, which might originate in material stripped from an extended, hydrogen-rich companion star. If our interpretation is correct, SN1991bg challenges both the double white-dwarf scenario, and the standard model for type Ia supernovae.

In the standard type Ia supernova model a massive carbon-oxygen white dwarf explodes, leaving no remnant. A mass of radioactive $^{56}\text{Ni} \sim 0.6 M_{\odot}$ (solar masses) is typically synthesized and ejected in the explosion⁶. This picture had already been challenged by determinations of a range $\sim 0.4\text{--}0.8 M_{\odot}$ of ^{56}Ni in different type Ia events (ref. 7). Serious objections to the double degenerate scheme for type Ia supernovae had also been raised from the negative results of searches for progenitor systems^{8–11}. These results have renewed interest in the single degenerate scheme¹², where a white dwarf grows to the point of explosive ignition by accreting matter from a nondegenerate companion. Such binaries might be symbiotic stars: interacting binaries with a late-type giant plus a hot compact companion (most often a white dwarf) which accretes material shed by the giant.

Are there enough massive white dwarfs in symbiotics to account for the type Ia supernova rate? It has been argued¹³ that the white dwarfs could accrete only $\sim 0.1\text{--}0.2 M_{\odot}$ before the system becomes a double white dwarf binary, since accretion from a stellar wind is only $\sim 10\%$ efficient and the companion has at most $\sim 1\text{--}2 M_{\odot}$ to lose. If only white dwarfs with masses very close to the Chandrasekhar mass explode, the number of suitable systems would be far too low because the observed white dwarf mass spectrum in symbiotics peaks below $0.6 M_{\odot}$. An alternative is that the progenitors might be carbon-oxygen (C+O) white dwarfs with mass $M \sim 0.5\text{--}1.3 M_{\odot}$ undergoing helium detonation at the surface after accreting $\sim 0.1\text{--}0.2 M_{\odot}$ of hydrogen-rich material, instead of central explosive carbon ignition close to the Chandrasekhar mass^{14,15}. Hydrogen would burn into helium, which would accumulate in a layer until explosive

ignition is triggered at the bottom of the layer just above the C+O interior. Two-dimensional hydrodynamic simulations¹⁶ demonstrate that a helium detonation initiated on the interface between the helium rich and the carbon-rich layers ignites a second detonation near the centre.

The hypothesis that the companion of the exploding white dwarf is a red giant or supergiant can be tested: Chugai¹⁷ predicted that in such a case the hydrogen-rich material, stripped from the extended companion by kinematic interaction with the supernova ejecta, would appear later as low-velocity material in

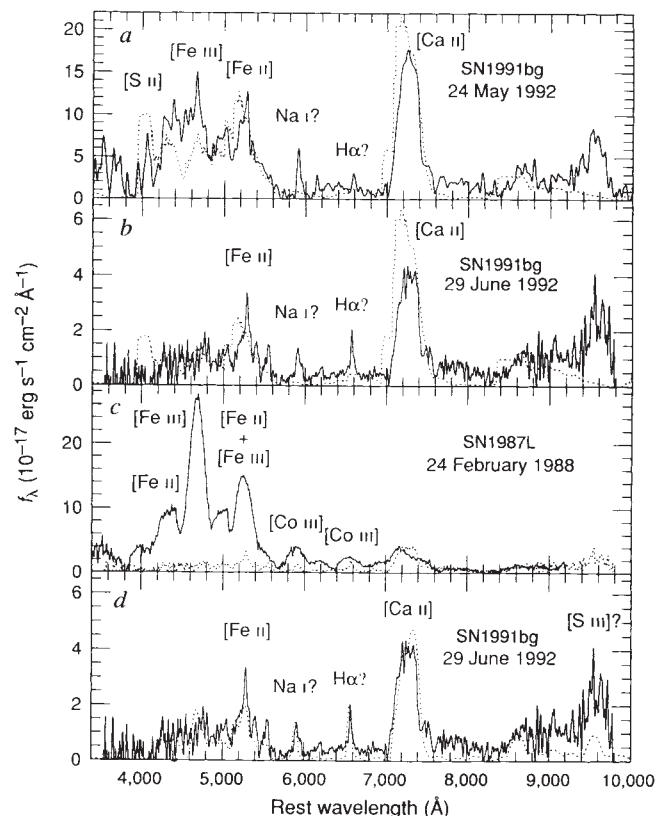


FIG. 1 a, The observed nebular spectrum of SN1991bg on 24 May 1992, taken with the 3-m Shane Telescope at Lick Observatory (solid line), and spectrum calculated from a model of a detonated C+O white dwarf of $0.65 M_{\odot}$ for day 161 after V maximum (dotted line). The V maximum is assumed to correspond to 15 days after explosion. This spectrum has been calculated with a new version of the code previously used to model the nebular spectra of SN1986G (ref. 25), adapted to deal with density and composition gradients. Given the uncertainties in our ionization treatments of calcium, we have reduced the amount of Ca II by a factor of ~ 5 within the iron-rich core, relative to that obtained in the model, both for this spectrum and for that of day 197 after V maximum. This reduction provides a better fit to the emission peak at around $7,300 \text{ Å}$. A reddening of $E(B-V) = 0.05$ mag has been used, based on the analysis of the interstellar Na I lines⁴. Note the different flux scales of the various plots. b, A nebular spectrum of SN1991bg obtained on 29 June 1992, taken with the 4.2-m William Herschel Telescope at La Palma Observatory (solid line), and a spectrum for day 197 after V maximum calculated as in a) (dotted line). c, Comparison of the observed spectrum of SN1991bg for day 197 after V maximum (dotted line) with the spectrum of a normal type Ia supernova (solid line) at about the same phase (~ 210 days after V maximum). The comparison spectrum is of SN1987L in NGC2336, obtained on 24 February 1988 at Lick Observatory with the 3-m reflector. d, A fit to the nebular spectrum of SN1991bg for day 197 after V maximum, based on an 'empirical' model (see text). Slow-moving H and Na I have been included to test the plausibility of the identifications of the narrow emission lines at $5,900 \text{ Å}$ and $6,570 \text{ Å}$.

TABLE 1 Explosion energies and abundances of detonated C+O white dwarfs

M^*	$E_{\text{expl}}^\dagger$	^{56}Ni	^{58}Ni	^{54}Fe	^{40}Ca	^{36}Ar	^{28}Si	^{32}S	^{24}Mg	^{20}Ne	^{16}O	^{12}C
0.60	0.46	0.012	0.001	0.011	0.002	0.013	0.206	0.097	0.011	0.002	0.231	0.011
0.65	0.57	0.055	0.001	0.014	0.004	0.015	0.231	0.109	0.010	0.002	0.203	0.009
0.70	0.69	0.133	0.001	0.015	0.004	0.015	0.229	0.109	0.008	0.001	0.176	0.007
0.75	0.82	0.234	0.001	0.014	0.004	0.014	0.214	0.102	0.007	0.001	0.151	0.006
0.80	0.95	0.342	0.001	0.013	0.004	0.013	0.194	0.093	0.006	0.001	0.127	0.005
0.85	1.07	0.450	0.001	0.012	0.004	0.011	0.172	0.082	0.005	0.001	0.107	0.004
0.90	1.19	0.558	0.001	0.010	0.003	0.010	0.150	0.072	0.004	0.001	0.087	0.003
1.00	1.42	0.764	0.001	0.008	0.003	0.007	0.106	0.051	0.002	0.000	0.056	0.002
1.20	1.83	1.122	0.001	0.003	0.001	0.002	0.036	0.017	0.001	0.000	0.016	0.000

Detonations are initiated by raising the temperature at the centre to induce thermonuclear runaway. Burning is first artificially propagated across a few layers at the speed of a Chapman–Jouguet detonation until it propagates self-consistently. Hydrodynamical equations are solved simultaneously with a network consisting of 13 nuclides²⁰. More detailed nucleosynthesis is calculated afterwards with a larger network. Those pure carbon–oxygen models are not entirely realistic, however, as an outer helium layer should be present for the proposed explosion mechanism to work. The mass of this layer would depend on the chemical composition of the material accreted¹⁴. If hydrogen is first accreted and burned into helium, H flashes might interact with the He layer below and, perhaps, induce He runaway burning for smaller He masses than in the case of direct He accretion.

* White dwarf masses and yields of the different nuclides are in $M_\odot$.

† In units of 10^{51} erg.

the nebular spectra of the supernova. His prediction has been reinforced by two-dimensional hydrodynamical simulations¹⁸.

The recent SN1991bg in NGC4374 (an elliptical galaxy in the Virgo cluster) provides evidence on both points: the mass of the white dwarf and the presence of a nondegenerate companion. SN1991bg was clearly underluminous at all observed epochs^{4,5}: at maximum light its blue (B) magnitude was ~ 2.5 mag fainter and its visual (V) magnitude ~ 1.6 mag fainter than a normal type Ia supernova in the same galaxy. Moreover, its light curve declined rapidly after maximum, as did that of SN1986G, a less subluminal type Ia (refs 4, 5, 19). Comparison of the $V-R$ colour at 30–80 days past maximum with those of other type Ia supernovae, and the weakness of the Na I D interstellar absorption line, favour low intrinsic luminosity over large interstellar extinction. The supernova entered the nebular phase sooner than other type Ia supernovae^{4,5}. This, together with the fast postmaximum decline of the light curve, points to a small mass of material above the Ni–Co core. The faint magnitudes indicate that only a small amount of ^{56}Ni was synthesized in the explosion, as this isotope powers the luminosity of type Ia supernovae through the radioactive decay chain $^{56}\text{Ni} \rightarrow ^{56}\text{Co} \rightarrow ^{56}\text{Fe}$.

We analysed the spectra of SN1991bg obtained 161 and 197 days after V maximum (24 May and 29 June 1992 respectively), when the supernova was well into the nebular phase. This phase reveals the results of explosive burning of the white dwarf, in particular the inner region of the ejecta (hereafter called the iron-rich core) where ^{56}Ni is the major component. These spectra (shown in Figs 1a and 1b) clearly differ from those of normal type Ia supernovae (Fig. 1c).

From our calculations the electron temperature of the iron-rich core appears lower than in normal type Ia supernovae at the same phase ($\sim 4,200$ K instead of $6,800$ K, in the 24 May spectrum). The ionization is also low, especially in the last spectrum, where [Fe II] emission at $5,300$ Å dominates, instead of [Fe III] emission at $4,600$ Å. The narrow iron lines indicate that the iron-rich nebular core, where ^{56}Ni was most abundant, extends only up to $\sim 4,000$ – $5,000$ km s^{−1}. These facts are consistent with a small ^{56}Ni mass synthesized in the explosion (<0.1 – $0.2 M_\odot$, according to this analysis), and also with a small total mass ejected, as previously suggested^{4,5}.

Another remarkable characteristic of these spectra is the apparent H α emission at low velocity. If the emission comes from the supernova ejecta, it would confirm the origin of SN1991bg as a white dwarf plus red giant companion. The emission is seen clearly only in the spectrum of 29 June. It might be present in the spectrum of 24 May as well, but at this epoch the emission is not far above the noise level.

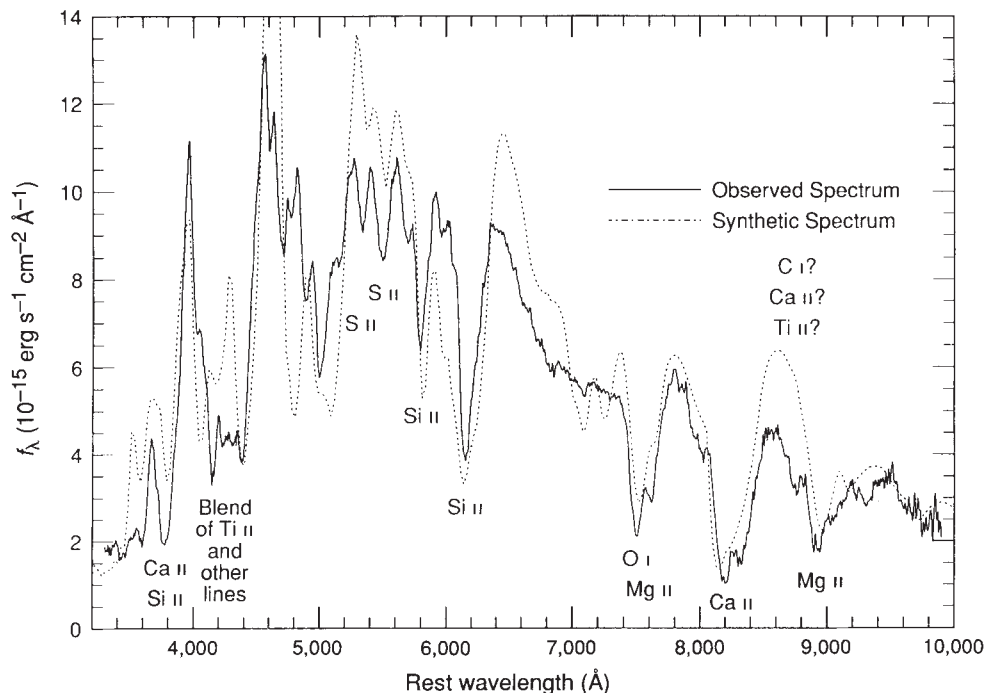
Figure 1d shows a calculated spectrum for day 197 after V maximum, based on an ‘empirical’ model (where the composition is chosen to best reproduce the nebular spectra). This model contains $0.06 M_\odot$ of ^{56}Ni in the inner emitting region, with velocities up to $4,000$ km s^{−1}. In this iron-rich core of the nebula, a region consisting mostly of hydrogen, with a small amount of sodium has been located moving at $1,000$ km s^{−1}. Intermediate-mass elements (Ca, S) have been included with velocities up to $7,000$ km s^{−1}. Given the uncertainty in the distance to the supernova and in the reddening (which are not investigated in this paper) an amount of initial ^{56}Ni between 0.05 and $0.1 M_\odot$ is determined from the good agreement between the observed and calculated spectrum.

To check the consistency of this composition with the idea that SN1991bg results from the detonation of a low-mass white dwarf, we have calculated (by means of a one-dimensional hydrodynamical model²⁰) the outcome of model detonations of C+O white dwarfs in the range $0.6 M_\odot \leq M \leq 1.2 M_\odot$. The explosion energies and the quantities of resulting nuclides are displayed in Table 1. Comparing this table with the analysis of the nebular spectrum on 29 June suggests that SN1991bg could result from the detonation of a white dwarf in the mass range 0.65 – $0.70 M_\odot$. This conclusion arises from the consistency of the nucleosynthesis obtained in these models with the results of the spectral analysis.

Calculation of the nebular spectra of the $0.65 M_\odot$ model at 161 and 197 days after V maximum are compared with the observed spectra of SN1991bg at the corresponding dates (Fig. 1a and 1b). This model accounts well for the emission in the spectra. This model synthesizes only $\sim 0.055 M_\odot$ of ^{56}Ni and $\sim 0.014 M_\odot$ of Fe (initially). The narrow features around $5,900$ Å (here identified as Na I D) and $6,570$ Å (here identified as H α) cannot be reproduced by the supernova model, which shows a broader velocity distribution. The narrow half-widths of these lines (~ 750 km s^{−1} for H α and $\sim 1,000$ km s^{−1} for the $5,900$ Å feature) suggest this material is low-velocity gas stripped from the companion during its interaction with the ejecta. The nuclear region of the host galaxy, NGC4374, is known to show H α and [N II] $6,548$, $6,583$ emission²¹, but the supernova was located far from the nuclear region and the inferred velocity from the half-width of the line (~ 750 km s^{−1}) is higher than typical velocities of the emitting gas in the galaxy (full-width at half-maximum of 350 km s^{−1})²². Also, the emission at $6,570$ Å is spatially aligned within 1 pixel (less than 1 arcsec) with the supernova emission in the CCD (charge-coupled device) two-dimensional spectral image.

To further test the $0.65 M_\odot$ white dwarf detonation model,

FIG. 2 Comparison of the spectrum of SN1991bg obtained one day before visual maximum, with that calculated from the detonated C+O white-dwarf model of $0.65 M_{\odot}$ (see text). Line identifications (tentative ones with a question mark) are indicated. One of the positive features of the model is its broad oxygen layer (at $8,000 \leq v \leq 22,500 \text{ km s}^{-1}$). Such a layer seems to be required to fit the spectra of all type Ia supernovae in the photospheric phases, not just SN1991bg (ref. 24). In the nebular phase this layer, containing $\sim 0.15 M_{\odot}$ of oxygen, does not give rise to significant forbidden emission as the dominant ionization stage is found to be O II, which is a less efficient coolant than O I. The ionization is thus higher here than in models of type Ib supernovae. An outer helium layer would have been burned to give iron-peak elements (mainly ^{56}Ni). An enhanced composition of initially synthesized ^{56}Ni in the low density outskirts of the supernova gives a characteristic signature in the premaximum spectra (this effect can be seen in SN1991T, refs 28, 29). This should no longer be visible at the phase when the first spectrum of SN1991bg was taken, as the photosphere would have already receded into the region where intermediate-mass elements are present. If a significant amount of unburned helium were left after the detonation of the helium layer, the spectrum may indicate the coexistence of He with ^{56}Ni (main product



we have calculated a model spectrum for ~ 1 day before V maximum using the local thermodynamic equilibrium approximation^{23,24} and, in Fig. 2, we compare it with the observed spectrum of SN1991bg from 14 December 1991⁴. The fit is reasonable.

We conclude that SN1991bg could have been produced by detonation of a low-mass ($M \approx 0.65 M_{\odot}$) C+O white dwarf accreting material from a nondegenerate hydrogen-rich companion. This result, combined with earlier results suggesting a range of ^{56}Ni production in type Ia supernovae (ref. 7), and the lack of sufficient progenitor systems for the double degenerate picture^{8,11}, challenge the standard massive white-dwarf model for type Ia supernovae and suggest that a mass range of detonated white dwarfs may account for these explosions. As Table 1 shows, detonations of C+O white dwarfs with masses $0.6 M_{\odot} \leq M \leq 1.0 M_{\odot}$ produce a range of ^{56}Ni from near zero up

to $\lesssim 0.8 M_{\odot}$. As we mentioned above, the smallest amount of ^{56}Ni previously estimated for a type Ia was $\sim 0.4 M_{\odot}$ (refs 7, 25), which still leaves a gap down to the $\sim 0.1 M_{\odot}$ of ^{56}Ni in SN1991bg. Another supernova which might have produced a small amount of ^{56}Ni was SN1971I (refs 4, 19, 26). If type Ia supernovae are produced by detonation of C+O white dwarfs filling a continuous range of masses, their use as standard candles is seriously compromised. It is important to determine the actual range and distribution of white-dwarf masses that explode as type Ia supernovae, calculate the luminosity of these explosions, and evaluate their detection probability. If C+O white dwarfs with masses $\lesssim 0.6 M_{\odot}$ detonate, they will generally go unnoticed, because they produce negligible amounts of ^{56}Ni . However, their contribution to the nucleosynthesis of intermediate-mass elements might be important as the white-dwarf mass spectrum peaks at $\sim 0.58 M_{\odot}$ (ref. 27).

of He incineration). Such coexistence, investigated in the context of a larger mass of helium and a different overall composition, gives rise to strong He I lines through nonthermal excitation by fast electrons arising from Compton scattering of γ -rays resulting from the decay of ^{56}Co (ref. 30).

Received 20 May; accepted 17 August 1993

1. Woosley, S. E. & Weaver, T. A. *Rev. Astr. Astrophys.* **24**, 205–253 (1986).
2. Iben, I., Jr., & Tutukov, A. V. *Astrophys. J. Suppl.* **54**, 335–372 (1984).
3. Branch, D. & Tammann, G. A. *Rev. Astr. Astrophys.* **30**, 359–389 (1992).
4. Filippenko, A. V. et al. *Astr. J.* **104**, 1543–1556 (1992).
5. Leibundgut, B. et al. *Astr. J.* **105**, 301–313 (1993).
6. Wheeler, J. C. & Harkness, R. P. *Rep. Prog. Phys.* **53**, 1467–1557 (1990).
7. Ruiz-Lapuente, P. & Filippenko, A. V. in *Origin and Evolution of the Elements* (eds Prantzos, N., Vangioni-Flam, E. & Cassé, M.) 318–322 (Cambridge Univ. Press, 1993).
8. Munari, U. & Renzini, A. *Astrophys. J.* **397**, L87–L90 (1992).
9. Robinson, E. L. & Shafter, A. W. *Astrophys. J.* **322**, 296–301 (1987).
10. Bragaglia, A., Greggio, L., Renzini, A. & D'Odorico, S. *Astrophys. J.* **365**, L13–L17 (1990).
11. Foss, D., Wade, R. A. & Green, R. F. *Astrophys. J.* **374**, 281–287 (1991).
12. Whelan, J. & Iben, I. Jr. *Astrophys. J.* **186**, 1007–1014 (1973).
13. Kenyon, S. J., Livio, M., Mikolajewska, J. & Tout, C. A. *Astrophys. J.* **407**, L81–L84 (1993).
14. Woosley, S. E. & Weaver, T. A. *Astrophys. J.* (submitted).
15. Livne, E. *Astrophys. J.* **354**, L53–L56 (1990).
16. Livne, E. & Glasner, A. S. *Astrophys. J.* **370**, 272–281 (1991).

17. Chugai, N. *Soviet Astr.* **30**, 563–567 (1986).
18. Livne, E., Tuchman, Y., Wheeler, J. C. *Astrophys. J.* **399**, 665–671 (1992).
19. Phillips, M. M. et al. *Publ. Astr. Soc. Pacif.* **99**, 592–605 (1987).
20. Canal, R., Isern, J., & Labay, J. *Astrophys. J.* **398**, L49–L52 (1992).
21. Burbidge, E. M. & Burbidge, G. R. *Astrophys. J.* **142**, 634–640 (1965).
22. Hansen, L., Nørgaard-Nielsen, H. U. & Jørgesen, H. E. *Astr. Astrophys.* **149**, 442–448 (1985).
23. Jeffery, D. J., et al. *Astrophys. J.* **397**, 304–328 (1992).
24. Kirshner, R. P. et al. *Astrophys. J.* (in the press).
25. Ruiz-Lapuente, P. & Lucy, L. B. *Astrophys. J.* **400**, 127–137 (1992).
26. Phillips, M. M. *Astrophys. J.* **413**, L105–L108 (1993).
27. Weidemann, V. & Koester, D. *Astr. Astrophys.* **132**, 195–202 (1984).
28. Filippenko, A. V. et al. *Astrophys. J.* **384**, L15–L18 (1992).
29. Ruiz-Lapuente, P., et al. *Astrophys. J.* **387**, L33–L36 (1992).
30. Lucy, L. B. *Astrophys. J.* **383**, 308–313 (1991).

ACKNOWLEDGEMENTS. The authors acknowledge the support of grants from NASA, NSF, ST Sci, DGICT and the DGU de Catalunya.

Split comets and the origin of crater chains on Ganymede and Callisto

H. J. Melosh* & P. Schenk†

* Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA

† Lunar and Planetary Institute, 3600 Bay Area Boulevard, Houston, Texas 77058, USA

WHEN the Voyager 1 spacecraft flew through the jovian system in January 1979, it returned images of several prominent chains of impact craters on the surface of the moon Callisto (Fig. 1). These impressively straight chains, or catenae, are composed of between 4 and 25 craters, and are up to 620 km long. They were initially thought to be secondary craters produced by debris from a larger primary impact¹, but detailed searches for source craters have been largely unsuccessful: a satisfactory explanation for the crater chains has yet to be found. Inspired by the recent observations of comet Shoemaker–Levy 9, which split into a line of about 20 fragments as it swept past Jupiter², we suggest that the impact of previous split comets might be responsible for at least some of the catenae on Callisto. In support of this hypothesis, we find that nearly all of Callisto's crater chains are on the Jupiter-facing hemisphere, as are an additional three catenae that we have found on Ganymede. We present a simple model of tidal breakup which both reproduces the range of observed chain lengths and indicates that the parent comets responsible for the Callisto catenae were typically no more than about 10 km in diameter.

If crater chains on Callisto are created by the impact of tidally split comets, a number of testable consequences immediately follow. First, because comets' or asteroids' orbits through the jovian system are generally hyperbolic (Shoemaker–Levy 9 is an exception), the impact of a fragment chain on one of Jupiter's tidally locked satellites should generally occur on the Jupiter-facing side. (Note that we cannot distinguish between comets and asteroids from the crater chains alone; in the following text we use the term 'comet' for either class of object.) Thirteen crater chains that are not obviously secondary to large basins have been recognized on the 70% of Callisto's surface that has been imaged. Of these, 11 are on the Jupiter-facing hemisphere and one is just over the edge on the opposite (anti-Jupiter) hemi-



FIG. 1 Illustration of a crater chain overprinting the rings of the Valhalla basin on Callisto. The chain is ~340 km long. Voyager image no. FDS 16424.32. This image was enhanced by a de-smearing algorithm written by M. Hicks.

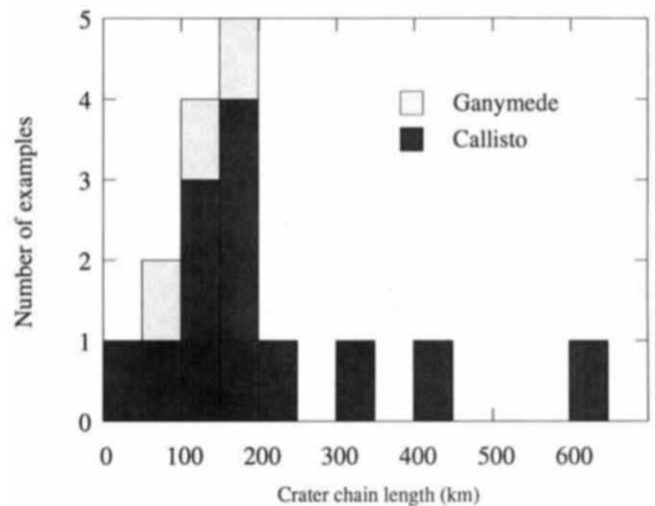


FIG. 2 Histogram showing the lengths and number of crater chains that are not obviously associated with primary craters on Ganymede and Callisto. Note that, on average, crater chains on Ganymede are shorter than those on Callisto, although statistics are poor.

sphere. The single chain that is near the centre of the anti-Jupiter hemisphere contains a number of highly elliptical craters and is morphologically distinct from all the other chains. Furthermore, if such split comets strike Callisto at its distance of $26.2 R_J$ from Jupiter (R_J is the radius of Jupiter, 71,400 km), some crater chains should also be visible on Ganymede, at a range of $14.9 R_J$. We have found three such chains on the Jupiter-facing hemisphere of Ganymede. Although a few crater chains also occur on the opposite hemisphere, these are secondary to the large craters Halieus and Gilgamesh. The smaller number of crater chains on Ganymede may be explained by its younger surface. At Ganymede's smaller range from Jupiter, however, we expect the chains to be shorter, as the fragments do not have as much time to separate after their parent is tidally disrupted inside Jupiter's Roche limit (the zone within which tidal forces tending to disrupt a small body exceed self-gravitational binding forces³, which lies at $2.67 R_J$ for comets of density 1 g cm^{-3}). Figure 2 shows the statistics of crater chain length on Ganymede and Callisto. Although the chains of Ganymede are small in number, they are on average shorter than the chains on Callisto.

We constructed a model of comet splitting, and investigated the separation of the fragments as a function of several parameters to investigate further the possibility that the crater chains on Ganymede and Callisto are created by split comets. These parameters include initial comet size, closeness of approach to Jupiter, and initial velocity of approach far from Jupiter. Comets rotate with periods long compared to the time of passage inside the Roche limit. For example, the period of comet Halley is now estimated⁴ as ~53 h (although with some uncertainty) whereas even slow (parabolic) comets spend at most ~1 h inside the Roche limit. Rotation can therefore be neglected for this problem.

The tidal stresses induced in either spherical⁵ or even ellipsoidal⁶ elastic bodies have tensional axes parallel to the radius vector. This stress state tends to separate the body along planes normal to the radius vector⁷. We therefore modelled the tidal disruption of a comet by assuming that it breaks into a number of fragments at the point of closest approach to Jupiter, where tidal stresses are greatest. This is an approximation, but it probably makes little difference if disruption actually occurs a little before or after closest approach. When the comet breaks up, we trace the orbits of the two extreme fragments that are broken off at $R_{\min} + r_{\text{comet}}$ and $R_{\min} - r_{\text{comet}}$, where $R_{\min}$ is the distance of the closest approach to Jupiter and r_{comet} is the radius

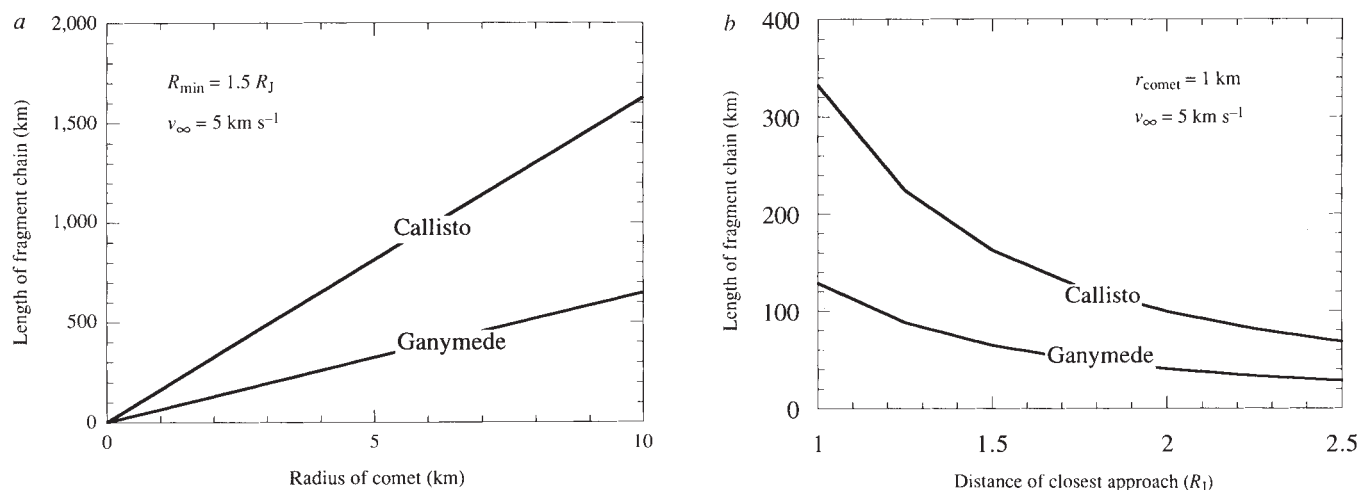


FIG. 3 Results of modelling the orbital evolution of a chain of fragments from a comet split by tides during a close approach to Jupiter. *a* Illustrates how the length of the fragment chain depends on the initial comet radius for impacts on either Ganymede or Callisto. $R_{\min}$ is the distance

of closest approach to Jupiter and v_{∞} is the approach velocity of the comet to Jupiter at great distance. *b* Illustrates how the fragment chain length depends on the distance of closest approach to Jupiter, in units of jovian radii, R_J for a comet of radius $r_{\text{comet}} = 1 \text{ km}$.

of the comet. The initial velocity assigned to the fragments is zero with respect to their mutual centre of mass; the entire difference between their subsequent orbital paths is due to their slightly different starting positions and, to a lesser extent, the approach velocity of their parent comet. Cometary rotation would actually give the different fragments slightly different initial velocities, but this effect is small compared with the other uncertainties. We followed the subsequent dispersion of the fragments using a fifth-order adaptable Runge-Kutta integrator⁸ for the centre of gravity of the fragments and for their relative separation. The equation of motion was expanded and linearized about the centre of mass, a sufficient approximation so long as the fragment separation is small compared to the distance to Jupiter, which was true in all cases studied.

The fragments from the split comet form a well defined line in space only for this model of radial breakup. We also constructed a model of rotational bursting of a comet in which tidal stresses are presumed to trigger instantaneous disintegration of a rotating comet, with the fragments flying off in the former equatorial plane with initial velocities determined by the speed of rotation. In this model the fragments formed an elongated loop, not a line. Although the aspect ratio of this loop is large at Callisto, often 6:1, the impact of such a loop would not produce the observed straight crater chains. The apparently linear dispersion of the fragments from comet Shoemaker-Levy 9 therefore suggests a radially separated breakup mode. Moreover, a detailed application of the radial breakup model to Shoemaker-Levy 9 gives excellent agreement between the predicted and observed locations of the fragments⁹.

Figure 3 shows the predicted lengths of a fragment chain at the orbits of Ganymede and Callisto for comets travelling directly from perijove to the satellite: fragment dispersion becomes so large after this first pass that fragments in orbit about Jupiter would not form continuous crater chains on the satellites. Figure 3*a* shows that the dispersion is a linear function of comet radius, and is about twice as large at Callisto as at Ganymede. Because the ~20–37 km diameter craters in Gupul catena were probably made by objects a few kilometres in diameter, and may thus have evolved from a parent comet ~5 km in diameter, the predicted chain length of 400 km in this figure is consistent with the observed 620 km length, given the variability in the angle at which the line of fragments strikes the surface and the distance of closest approach to Jupiter (assumed to be $1.5 R_J$ in Fig. 3*a*). Figure 3*b* shows the dependence of fragment chain length on distance of closest approach to

Jupiter for a comet of radius 1 km (comparable to the currently estimated dimensions⁹ of Shoemaker-Levy 9). The predicted crater chain lengths are comparable to those observed, suggesting that most such chains are made by comets in this size range. This figure indicates that the dispersion is inversely proportional to encounter distance, but that the dispersion at Ganymede is about half that on Callisto under similar conditions. Varying the velocity of approach shows that the dependence of chain length on this factor is weak (~30% decrease) when approach velocities are varied from 0 to ~10 km s⁻¹. The dispersion then decreases by about a factor of two at 20 km s⁻¹ (the maximum Jupiter approach velocity for comets bound to the sun is 31 km s⁻¹). The dependence of chain length on approach velocity is thus not as strong as the dependence on initial comet size or distance of closest approach to Jupiter.

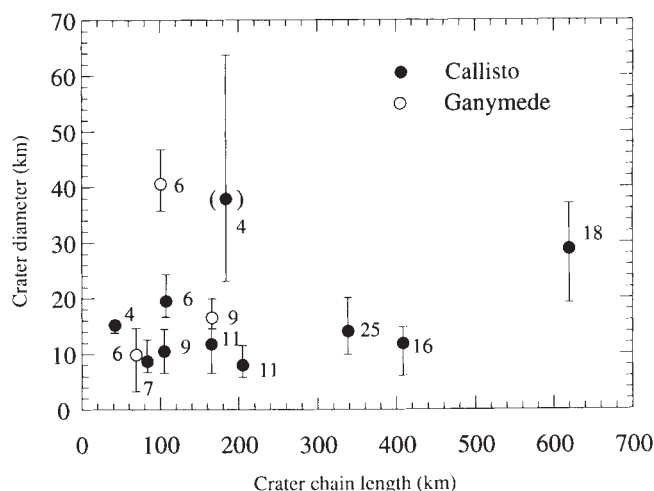


FIG. 4 Correlation between crater chain length and crater diameter for crater chains on Ganymede (open circles with points) and Callisto (solid circles). 'Error bars' indicate the range between maximum and minimum crater sizes; the symbol indicates the average. The number next to each symbol indicates the number of craters in the chain. The data point surrounded by parentheses is not on the Jupiter-facing hemisphere of Callisto and thus may be secondary to some unseen basin.

Assuming that the size of the fragments scale with the size of the parent bodies, the simple model presented here predicts that there should be a positive correlation between crater chain length and crater size. (Considerable scatter should occur because of varying angles of encounter between the fragment chain and satellite surface, encounter distance, velocity of approach, and the number of fragments formed.) A very rough correlation between crater chain length and average crater diameter per chain on Ganymede and Callisto may be present when chains with similar numbers of craters are compared (Fig. 4: note that only ten crater chains on Callisto are plotted because poor image quality prevents accurate diameter measurements on three chains). For a given chain length, there is also a tendency for average crater diameter to increase as the number of craters (and hence comet fragments) per chain decreases, as expected from volume conservation. The order of magnitude of the dispersion is also consistent with the comet splitting model.

On the other hand, this correlation depends in large part on the single data point for the 620-km-long Gipul catena. An alternative interpretation of Fig. 4 is that the craters of nearly all the chains in Fig. 4 are ~10 km in diameter, with the exception of the craters in Gipul. In this interpretation the parent comets are viewed as loose aggregations of pre-existing fragments that are all roughly the same size. Such loose aggregations would then be dispersed by the weak tidal forces during passage inside the Roche limit, as has been suggested⁹ for comet Shoemaker–Levy 9. Small parent comets would thus contain only a few such fragments, and large parents would contain many. Gipul's parent must have contained exceptionally large fragments.

If crater chains on Ganymede and Callisto are created by tidally split comets, they preserve a fossil time record of the occurrence and characteristics of comet fragmentation. We esti-

mate the recurrence interval between comet-splitting events from the number of crater chains that have accumulated on the surface of Callisto, assuming a constant cratering rate since the surface formed ~4.0 Gyr ago. Voyager produced adequate images of 70% of the Jupiter-facing hemisphere of Callisto. If it is assumed that the observed 12 crater chains were all produced by split comets, and that comets are equally likely to approach Jupiter from all directions, then the following relation may be deduced. The number of comets that have split near Jupiter is larger than the number on the surface of Callisto by the ratio between the area of a sphere of radius equal to the radius of Callisto's orbit and the projected surface area of Callisto (neglecting gravitational focusing). This relation yields an estimate of $\sim 4.9 \times 10^7$ split comets over a period of 4 Gyr, or a recurrence interval of ~80 yr between splitting events, which is consistent with the observation of the splitting of comet Brooks 2 near Jupiter in 1886 (ref. 10). $\square$

Received 5 August; accepted 30 September 1993.

1. Passey, Q. R. & Shoemaker, E. M. in *Satellites of Jupiter* (eds Morrison, D. & Matthews, M. S.) 379–434 (Univ. of Arizona Press, Tucson, 1982).
2. Chapman, C. R. *Nature* **363**, 492–493 (1993).
3. Stacey, F. D. *Physics of the Earth* 313–317 (Wiley, New York, 1977).
4. Belton, M. J. S. in *Comets in the Post-Halley Era* (eds Newburn, R. L. J., Neugebauer, M. & Rahe, J.) 691–721 (Kluwer, Dordrecht, 1991).
5. Sekiguchi, N. *The Moon* **1**, 429–439 (1970).
6. Dobrovolskis, A. R. *Icarus* **52**, 136–148 (1982).
7. Dobrovolskis, A. R. *Icarus* **88**, 24–38 (1990).
8. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, 1986).
9. Scotti, J. V. & Melosh, H. J. *Nature* **365**, 733–735 (1993).
10. Sekanina, Z. & Yeomans, D. K. *Ast. J.* **90**, 2335–2352 (1985).

ACKNOWLEDGEMENTS. Lunar and Planetary Institute contribution no. 817.

Estimate of the size of comet Shoemaker–Levy 9 from a tidal breakup model

J. V. Scotti & H. J. Melosh

Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA

THE spectacular 'string-of-pearls' appearance of periodic comet Shoemaker–Levy 9 (1993e) and its impending encounter with Jupiter's atmosphere have caused a flurry of speculation on the likely effects of the impact¹. The magnitudes of the predicted effects depend critically on the masses of the fragments composing the chain, although these are currently poorly constrained. However, some limits on the comet's total size, and hence mass, can be obtained by considering its dynamical history. On its previous swing past Jupiter, the comet apparently passed within the Roche limit of the planet^{2,3}, and it therefore seems likely that the present chain of about twenty nuclei was created when a single progenitor was split by tidal forces during closest approach. Here we describe a simple model of tidal splitting which is able to reproduce closely both the length of the observed chain and its position angle on the sky as a function of time. The length of the fragment chain is directly proportional to the size of the parent object, which we estimate to have been only about 2 km in diameter.

Soon after the discovery of Shoemaker–Levy 9, orbit calculations by Marsden³ and Nakano² (which used measurements of the midpoint of the fragment chain) indicated that the object is now a temporary satellite of Jupiter and had passed within the Roche limit on or about 8 July 1992. A second encounter, which includes an impact of the object on Jupiter, is predicted in late July 1994. The position of the comet relative to Jupiter is shown

TABLE 1 Length and orientation of the fragment train

Date	l_{obs} (arcsec)	l_{pred} (arcsec)	l (10^3 km)	θ_{obs} (degrees)	θ_{pred} (degrees)	ϕ_{full} (degrees)
1993 March 26.3	49	50	164	76.0	77.5	20.3
1993 March 28.3	53	51	166	76.0	77.5	20.5
1993 March 30.3	51	52	168	76.6	77.5	20.7
1993 April 20.3	54	57	189	77.3	77.0	23.5
1993 May 17.2	61	62	218	74.7	76.5	29.9
1993 May 27.2	68	64	229	75.3	76.4	33.8
1993 June 26.2	69	68	266	75.5	76.0	59.1

The observed length of the fragment train is l_{obs} , l_{pred} is the predicted length for an extreme fragment centre at radius $r_c = 0.815$ km, l is the full three-dimensional (not projected) length of the train, θ_{obs} is the observed position angle of the train, θ_{pred} is the predicted position angle and ϕ_{full} is the full (not projected) angle between the fragment train and the joventric velocity vector.

in Fig. 1 for the period between closest approach last year and the impact next year.

The close approach to Jupiter in July 1992, combined with the uncertainties in the orbit determination of Shoemaker–Levy 9 do not allow us to determine accurately the past dynamical history of the comet. We can, however, integrate the orbit with an ensemble of slightly different orbits back past the 1992 encounter with Jupiter to get an idea of how this comet came to be on its present orbit. We find that it is very likely that the comet was captured into orbit around Jupiter in 1971. Nearly half the orbits, however, remained bound to Jupiter until the end of the simulation, 80 years ago.

CCD (charge-coupled device) observations of the object using the 91-cm Spacewatch telescope on Kitt Peak in Arizona have been obtained and analysed to measure the relative positions of the brightest nuclei and the length and orientation of the chain. It should be noted that the fragments are not dispersed along the orbit itself: they form an apparently straight line oriented at a large angle to the orbital direction (see Table 1, where the

position angle projected on the sky and the true angle between the fragment chain and the jovian orbit are recorded for seven observations from March to June).

A simple model has been constructed for tidal splitting of comets or asteroids⁴ and it was shown that the crater chains on Ganymede and Callisto may have been created by the impact of such objects. Although this model was created for objects on hyperbolic orbits with respect to Jupiter, it is straightforward to apply the same model to Shoemaker–Levy 9. We assume that the object split into a number of fragments at the point of closest approach to Jupiter, where tensile tidal stresses exceeded self gravity (this occurs within the Roche limit, which lies $\sim 2.7 R_J$ from the centre of Jupiter, where R_J is the radius of Jupiter). Models of tidal disruption of solid bodies⁵ show that the initial (tensile) fracture occurs along planes perpendicular to the axis of maximum tension, which is along the line connecting the centres of the comet and Jupiter. After fracture, we assume that the fragments travel independently of one another (we neglect mutual gravity, whose effect we believe is small). The orbits of the extreme members of the chain are constructed by adding (and subtracting) the radius of the object, r_c , to (or from) the distance from Jupiter of the parent object's centre at the time of closest approach. Rotation of the parent is neglected, so no initial velocities are added. The subsequent dispersion of the fragments is thus entirely due to their slightly different starting points during closest approach.

The mean orbit of Shoemaker–Levy 9 was taken from ref. 6, and integrated in a model Solar System that includes the nine planets Mercury to Pluto using a fifteenth-order RADAU integrator⁷. The chain length and position at any given time of observation were constructed by adding a trial radius, $\pm r_c$, to the mean position at perijove, then integrating to the desired time. Note that r_c is thus the distance from the centre of the parent comet to the centre of the extreme fragment. The parent's actual radius is therefore larger than r_c . The predicted chain length was compared to the observations (the orientation is independent of the radius; it is completely determined by the breakup model), and the radius adjusted until a best-fit radius is achieved. The results of this process are reported in Table 1,

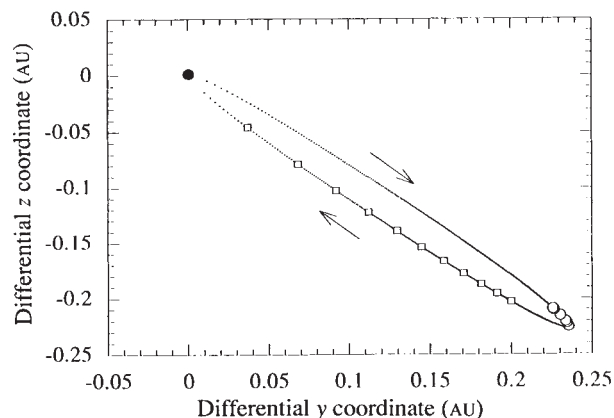


FIG. 1 The position of comet Shoemaker–Levy 9 plotted with respect to Jupiter in ecliptic (1950) coordinates. Coordinate z is the height above the ecliptic plane and the orbit is projected on the y - z plane. Jupiter's location is indicated by the large black dot at (0, 0), which is many times larger than Jupiter itself at this scale. Each small dot represents the daily location of the chain's midpoint beginning on 10 July 1992 and ending on 18 July 1994. The circled points show the location of the midpoint on the days listed in Table 1, and the squares mark its location on the dates listed in Table 2. The length of the chain is too short to show to scale except for the months just before impact. (AU indicates astronomical units; 1 AU is the mean radius of the Earth's orbit).

TABLE 2 Predictions for the length and orientation of the train in 1994

Date	l_{pred} (arcmin)	l (10^3 km)	θ_{pred} (degrees)	ϕ_{full} (degrees)
1993 Dec 19	2.06	566	66	1.59
1994 Jan 8	2.33	616	65	2.42
1994 Jan 28	2.66	673	63	2.84
1994 Feb 17	3.08	736	62	3.01
1994 Mar 9	3.62	810	62	3.02
1994 Mar 29	4.24	898	62	2.87
1994 Apr 18	4.93	1,005	63	2.58
1994 May 8	5.69	1,144	64	2.17
1994 May 28	6.57	1,337	64	1.65
1994 June 17	7.79	1,645	64	1.08
1994 July 7	10.61	2,365	63	0.99

The predicted length of the chain is l_{pred} , l is the full three-dimensional (not projected) length of the train, θ_{pred} is the predicted position angle and ϕ_{full} is the full (not projected) angle between the train and the jovian orbital velocity vector.

in which the predicted and observed orientations and lengths of the chain are in excellent agreement. The slight misfit in angle of $\sim 1^\circ$ may be a result of neglecting initial rotation. Nevertheless, the overall agreement shows that the elementary model captures most of the physics of the disruption process. The lengths are reported in arc seconds subtended on the sky, and the position angle is in degrees measured anticlockwise from north through east in equatorial coordinates. The best-fit initial separation of the extreme fragments, $2r_c$, is ~ 1.6 km, using the mean orbit's⁶ closest approach to Jupiter of $1.57 R_J$. This separation estimate is proportional to the approach distance, and varies $\sim 10\%$ for variations of $0.1 R_J$ around the assumed value. The actual parent comet was larger, as the calculated value gives the extreme range between the centres of the fragments, neglecting the radii of the fragments themselves. Also, the extreme antipodal regions of the parent may have been small, or perhaps composed entirely of deep regolith that would contribute background dust but no observable fragments. Assuming a 2.0-km extreme fragment separation for a progenitor that broke up into 20 equal-sized fragments, the mean fragment diameter is 0.74 km, and the parent comet diameter is 2.3 km.

Some constraints on the physical nature of the nucleus can be derived from an estimate of the tidal stress that acted on the comet during close approach to Jupiter. The problem of tidal stresses on a small body has been the subject of much sophisticated analysis⁵. An adequate approximation can however, be derived (up to constants of order one) by calculating the differential gravitational force on either half of the projectile, which are separated by nearly one projectile radius, then dividing by the cross-sectional area to give stress. This shows that the stress σ_{tide} is given by

$$\sigma_{\text{tide}} \approx 0.15 \frac{\rho G M_J}{R^3} r_c^2$$

where R is the distance to Jupiter's centre at closest approach, ρ is the density of the comet, G is the gravitational constant and M_J is the mass of Jupiter. Inserting a closest-approach distance of $1.5 R_J$ yields $\sigma_{\text{tide}} \approx 1.4 \times 10^{-4} r_c^2$, where σ_{tide} is in bar and r_c in km. The stress that apparently ripped Shoemaker–Levy 9 into ~ 20 fragments was thus only a fraction of a bar, suggesting that the parent nucleus had negligible strength. Perhaps a good model for the parent would be an incoherent aggregation of larger, pre-existing fragments and dust bound together entirely by gravitational forces.

The complex of debris surrounding Shoemaker–Levy 9 includes dust trails extending from the ends of the visible fragment chain. No additional nuclei can be resolved within these trails. This dust may be the remnants of a dust coma (existing before the Jupiter closest approach) which was also tidally dis-

persed. Alternatively, it may represent dust liberated at the time of closest approach and then perturbed by solar light pressure or other non-gravitational forces. These dust trails extended at least 17 arcmin end-to-end at the time of discovery. This corresponds to a debris complex with a diameter of ~ 20 – 35 times the diameter of the parent object at the time of closest approach to Jupiter, assuming that it was tidally dispersed.

In view of the success of the tidal splitting model in predicting the observed configuration of the fragment chain, we have run it forward in time to predict the length and orientation of the chain up to the time of impact on Jupiter. These predictions are given in Table 2. Using the best-fit extreme fragment separation of 1.6 km, we find that all the fragments will strike Jupiter and that the impacts will be spread over an interval of 5.6 days centred on 21 July 1994. Note that by the time of the impact the fragments are all nearly lined up along the orbit, reflecting the prediction that they will all strike the planet in nearly the same place relative to the dawn terminator. $\square$

Received 5 August; accepted 22 September 1993.

1. Chapman, C. R. *Nature* **363**, 492–493 (1993).
2. Nakano, S. *IAU Circ.* No. 5800.
3. Marsden, B. G. *IAU Circ.* No. 5800.
4. Melosh, H. J. & Schenk, P. *Nature* **365**, 731–733 (1993).
5. Dobrovolskis, A. R. *Icarus* **88**, 24–38 (1990).
6. Marsden, B. G. *Minor Planet Circ.* No. 22383.
7. Everhart, E. in *Dynamics of Comets: Their Origin and Evolution* (eds Carusi, A. & Valsecchi, G. B.) 185–202 (Reidel, Dordrecht, 1985).

Crystalline networks with unusual predicted mechanical and thermal properties

Ray H. Baughman* & Douglas S. Galvão*†

* AlliedSignal Inc., Research and Technology, Morristown, New Jersey 07962–1021, USA

† Instituto de Física, Unicamp, CEP 13081, Campinas, São Paulo, Brazil

MOST materials shrink laterally and become less dense when stretched. Materials that both expand laterally (that is, have negative Poisson's ratio) and densify when stretched are of interest both from the fundamental and the practical points of view^{1–5}. A few monocrystalline phases with negative Poisson's ratio are known^{3,4}, but these do not densify when stretched. Here we present molecular-mechanics calculations for some hypothetical phases of carbon which exhibit both kinds of behaviour. The properties derive from the presence of bonds that act as hinges in extended helical chains. Other unusual properties of these phases include negative thermal expansion, dopant-controlled porosity and low-temperature polymorphism. Such structures can be envisaged for polyacetylene, polydiacetylene, polyphenylene and (BN)_x phases, as well as for variants of some known, structurally related inorganic phases.

The phases described here are named twisted-chain auxetics, because phases with a negative Poisson's ratio are called auxetics¹ and because the auxetic property is the result of a new mechanism—the change in twist of helical chains. These helical chains result from the covalent interconnection of bonds from two sets of non-parallel 'basis' chains, which occupy alternating layers along the helical chain axis. Parallel bonds in the helical chains provide 'hinges' which enable a soft shear deformation mode. These hinge bonds can be, for example, sp^2 – sp^2 single bonds in polyacetylene chains. Tensile stress in the helical chain direction untwists these chains, thereby changing the interfacial angles of parallelepiped rods, whose corners correspond to four

parallel helical chains. For structures in which the shear deformation mode dominates, this process results in a negative Poisson's ratio in one across-diagonal interchain direction for the parallelepiped rods; a positive Poisson's ratio for the second such direction; and, in some cases, stretch-densification. Such a mechanism for achieving an auxetic material differs fundamentally from both the re-entrant-structure approach which forms the basis for auxetic foams and honeycombs, and an auxetic hydrocarbon phase proposed by Evans *et al.*^{1,3,5}. Mechanical models for achieving the auxetic property using rotatable nodes and attaching ligaments have also been proposed,³ again using quite different mechanisms.

The present work demonstrates that the shear deformation mode described above can result in negative volumetric thermal expansion coefficients. Negative volumetric thermal expansion is rarely observed, especially at high temperatures. No examples are available of three-connected network structures with this property, and a mechanistic relationship between negative thermal expansion and auxetic behaviour has not previously been established.

The twisted-chain auxetics can be divided into two extreme categories, 'crowded' and 'uncrowded', that have different properties. This division depends on whether or not the hinge deformation mode is severely constrained by repulsive interchain interactions, due to van der Waals overlap, that force the structure to a fully opened position. Only the uncrowded materials will show a special type of polymorphism, corresponding to two phases that are related by the hinge deformation mode. Also, one of these phases for the uncrowded twisted-chain auxetics is potentially stretch-densified for stress along the helical-axis direction, whereas severely crowded twisted-chain auxetics are stretch-rarefied along this direction.

Crystal structure and properties were calculated by molecular mechanics using the POLYGRAF program⁶ and an empirical force field^{7,8} which includes van der Waals, bond stretch, bond

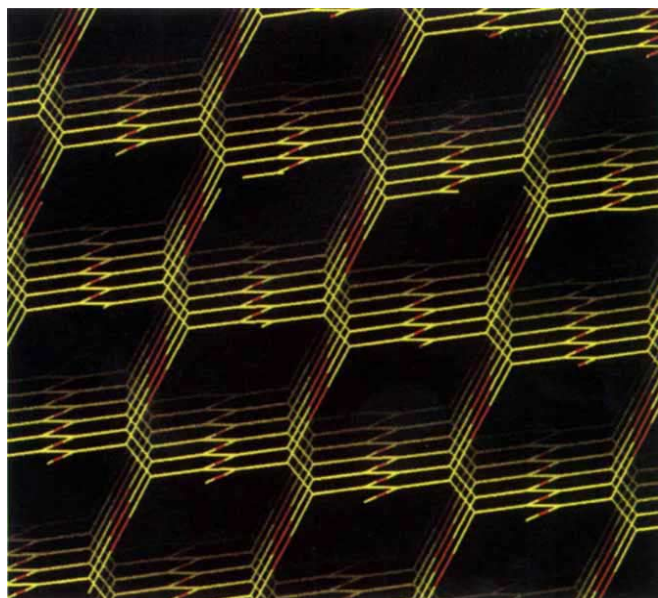


FIG. 1 The proposed crystal structure of the *trans* hinged polydiacetylene carbon, viewed normal to the hinge bonds that connect the two sets of non-parallel polydiacetylene basis chains. For clarity, the $C\equiv C$ bonds in the polydiacetylene chains of the carbon network are indicated using red lines. The helical polyacetylene chain axis, which provides the auxetic behaviour, passes through the centres of $C=C$ bonds from different polydiacetylene chains that are interconnected via a hinge bond.

angle bend and torsional rotation terms. Bond stretch and torsional potentials for conjugated phases were calculated from observed bond lengths^{9,10} using Allinger's relationships⁸ between π -bonding order and bond length, and between force-field potential and π -bonding order.

In order to explore the properties of an uncrowded twisted-chain auxetic, we have designed a phase that is particularly amenable to calculations, as it contains only two independent atoms. We call this phase the hinged polydiacetylene carbon, because it is a carbon phase analogue of polydiacetylene single crystals, which have been the focus of concentrated research since their discovery by Wegner¹¹ over twenty years ago. As shown in Fig. 1, all carbon atoms in the hinged polydiacetylene phase are in polydiacetylene basis chains; C=C bonds in neighbouring layers of such chains are connected through C—C hinge bonds to form helical polyacetylene chains. The non-equilibrium structure having maximum volume can deform via oppositely directed shear processes about the hinge bonds to generate two equilibrium phases. These are called the *cis* and the *trans* hinged polydiacetylene phases, according to the conformation distantly approached by the helical polyacetylene chains. Although no intracell symmetry was imposed for the molecular mechanics calculations, the derived structures (Table 1) for both these phases have the highest symmetry (*C2/c*) compatible with the specified covalent connectivity. The *trans* phase is only 0.17 kcal mol⁻¹ C lower in calculated energy than the *cis* phase, and the barrier for *cis*- to *trans*-phase conversion is only 0.14 kcal mol⁻¹ C, so facile interconversion between these phases is expected as

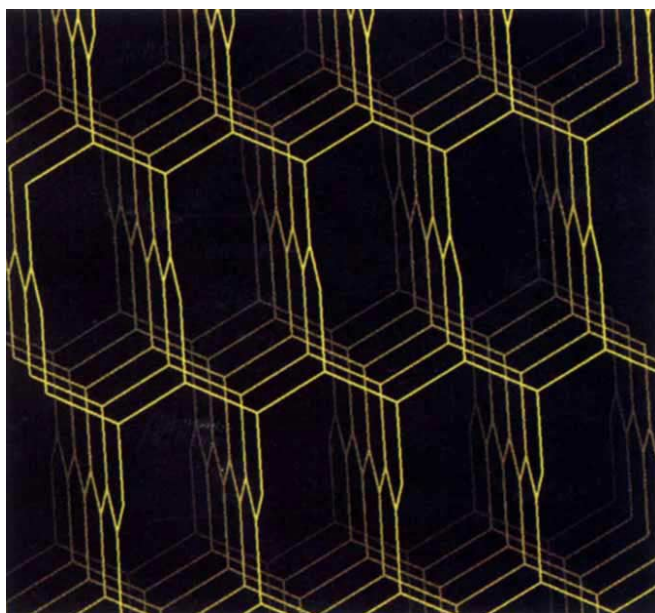


FIG. 2 The proposed crystal structure of the hinged polyacetylene carbon, viewed approximately parallel to one set of polyacetylene chains and approximately orthogonal to the second set. This crystal phase is crowded because interchain interatomic distances (2.5 Å) are much shorter than the sum of van der Waals radii. The resulting repulsive interactions force the structure to a fully opened arrangement, and eliminate the bandgap that is present for the known (CH)_x polyacetylene^{11,12}. Thermally excited rotations about the hinge bonds (which are approximately vertical in the figure) decrease volume, thereby providing the negative thermal expansion. The *c*-axis of the tetragonal unit cell for this phase is parallel to the hinge bonds. The auxetic behaviour is, however, best represented using a monoclinic unit cell (*C2/c*), with *c*-axis parallel to one of the four equivalent sets of helical polyacetylene chains. This chain axis coincides with a vector interconnecting the centres of backbone bonds in different *trans* polyacetylene basis chains that share the same hinge bond.

TABLE 1 Predicted structure and properties of carbon phases

	<i>Cis</i> hinged polydiacetylene	<i>Trans</i> hinged polydiacetylene	Hinged polyacetylene
Density (g cm ⁻³)	1.564	1.789	2.869
Cell parameters <i>a</i> (Å), <i>b</i> (Å), <i>c</i> (Å) Angle β	5.54, 8.06, 4.61 82.4°	8.81, 4.21, 4.88 99.5°	3.60, 3.60, 4.65 112.8°
Young's moduli <i>Y</i> ₁ , <i>Y</i> ₂ , <i>Y</i> ₃ (GPa)*	4.3, 18.9, 134	96, 4.5, 170	93, 75, 444
Poisson ratios*			
<i>v</i> ₁₂ , <i>v</i> ₁₃ , <i>v</i> ₂₃	0.45, 0.03, -0.02	4.1, -0.16, 0.07	0.94, -0.04, 0.16
<i>v</i> ₂₁ , <i>v</i> ₃₁ , <i>v</i> ₃₂	2.0, 0.84, -0.14	0.19, -0.28, 2.6	0.76, -0.18, 0.95
Atom coordinates†			
<i>x</i> ₁ , <i>x</i> ₂	0.9629, 0.8121	0.9640, 0.8135	0.9132
<i>y</i> ₁ , <i>y</i> ₂	0.0597, 0.1935	0.0638, 0.1958	0.1250
<i>z</i> ₁ , <i>z</i> ₂	0.1010, 0.0300	0.1010, 0.0297	0.0764

So that direct comparison can be made between the results for these different phases, the hinged polyacetylene phase is represented using a monoclinic unit cell which is defined analogously to that for the hinged polydiacetylene phases.

* Mechanical properties are with respect to orthogonal axes which coincide with the *b* and *c* crystallographic axes. The Poisson's ratios (*v*_{*ij*}) are the ratios of the contractile strains for lateral dimensions (ϵ_j) to the elongational strains for the tensile stress direction (ϵ_i).

† Subscripts 1 and 2 denote fractional coordinates for *sp*² and *sp* carbon atoms, respectively.

a function of mechanical stress and temperature. Because the *trans* phase has a much higher density (1.789 g cm⁻³) than the *cis* phase (1.564 g cm⁻³) and lower vibrational freedom, high temperatures and low pressures favour conversion to the *cis* phase. As shown in Table 1, both of these phases have two negative Poisson's ratios, for stress either along the helical axis (*c*-axis) or one orthogonal direction. Additionally, the *cis* and *trans* phases are stretch-densified for stress along the *b*-axis and along both the *a*-axis and *c*-axis, respectively—because the sum of the two Poisson's ratios orthogonal to these individual directions exceed unity.

A carbon phase consisting of polyacetylene basis chains provides a prototypical example of a severely crowded twisted-chain auxetic. This phase (shown in Fig. 2), which is referred to herein as the hinged polyacetylene carbon, is isomorphous with many known phases¹², and has been predicted to be a tetragonal metal^{13,14}. We use a monoclinic unit cell to represent mechanical properties, so that a helical polyacetylene chain formed by the interconnection of planar polyacetylene chains becomes an axial direction (*c*-axis). In contrast to the discussed case of an uncrowded twisted-chain auxetic, only one equilibrium structure exists at zero pressure, and the helical axis does not correspond to an axis of stretch-densification. The absence of both of these features is a consequence of major overlap of van der Waals volumes, which provides an increased energy for the hinge deformation mode compared with other deformation modes. In fact, if the van der Waals repulsion is increased, such as by decreasing bond lengths, even the auxetic behaviour disappears. Although the auxetic behaviour was not previously recognized, a bulk modulus (calculated *ab initio*) is available¹⁴ for the hinged polyacetylene phase (362 GPa). The similarity between this value, and the value obtained in the present work (382 GPa) supports the validity of the molecular mechanics calculations.

Doping processes can convert an uncrowded twisted-chain auxetic to a crowded twisted-chain auxetic. For example, the electron affinity of polydiacetylene chains¹⁵ is high enough to ensure that doping the hinged polydiacetylene phases with alkali metals will be highly exothermic. Such doping is expected to result in a transition from a large bandgap semiconductor to a metal. Doping with large radius ions would, however, substantially increase the effective modulus for the hinge deformation

mode, thereby eliminating both the stretch-densification property for the helical-axis direction and the special type of polymorphism.

In extending the family of twisted-chain auxetics to other chemical compositions, it is important to recognize that phases that are topologically related to the phases described above are not necessarily auxetic. For example, hinged polyethylene phases having only positive Poisson's ratios can be obtained by adding hydrogen to the sp^2 and sp carbon atoms in the hinged polydiacetylene phases, producing chains with the polyethylene backbone. Auxetic behaviour is absent because overlapping van der Waals volumes constrain the hinge deformation mode to the point that other deformation mechanisms dominate. Nevertheless, an infinite variety of twisted-chain auxetics can be derived that are based on the same three-dimensional, three-connected network as the hinged polydiacetylene and hinged polyacetylene phases. Various alternative basis-chain types can be employed, ranging from hydrocarbon chains based on polyphenyls to BN chains. Generally, these basis chains can be either regular or non-periodic copolymers, alternate layers of basis chains can differ, and different layers of hinge elements can be independently expanded—for example, by replacing the C—C hinge bonds by either C—C≡C—C or *para*-connected phenyls. Also, the twisted-chain auxetic need not have covalent bonding that is extended in three dimensions, because the operative parallelepiped elements can be incorporated in two-dimensional sheet structures.

The soft shear mode of the twisted-chain auxetics can provide another unusual property—negative volumetric thermal expansion. This aspect is easiest to understand for the hinged polyacetylene phase shown in Fig. 2, as oppositely directed shear deformations about the hinge bonds provide identical volume decreases. Molecular dynamics calculations of average cell parameters as a function of temperature (using POLYGRAF⁶ and the above described force field) provide a thermal expansion coefficient at room temperature of $-1.6 \times 10^{-5} \text{ K}^{-1}$ for this phase. In addition to the sign, the magnitude of this coefficient is unusual, as it is about an order of magnitude larger than is typical of materials with covalent structures extended in three dimensions. For this and other severely crowded structures, the magnitude of the calculated negative thermal expansion coefficient is decreased by even small changes in bond lengths that increase crowding.

The presently proposed twisted-chain auxetics may have interesting applications. Such materials would provide a high Young's modulus for stress along an axis that provides auxetic behaviour. This property, which is desirable for some types of engineering applications, is not obtainable for materials that are auxetic because of pore structure³. Also, the shear deformation mode, which is inherent to these structures, would provide materials toughness and damage resistance that are atypical for structures having three-dimensional covalent bonding. For structures that are conjugated, such as the hinged polydiacetylene phases, interesting properties might be anticipated by analogy with the observed properties of polydiacetylene crystals, such as piezochromism, electrochromism, thermochromism, and photochromism^{16,17}. Uncrowded phases might also be useful for redox device applications, such as displays or electromechanical actuators¹⁸, because the shear deformation mode provides a mechanism for rapid incorporation of various size dopants, and covalent connectivity in three dimensions provides the possibility of cycle lifetimes that are unobtainable for ordinary conjugated polymers.

Because a large variety of materials are twisted-chain auxetics, a host of synthetic routes are available for exploration—such as those described by Diederich and Rubin¹⁹ for other forms of carbon. Some members of this family might be already synthesized, but not yet identified as having unusual properties. Possible candidates are the various known phases of the ThSi_2 or LaPtSi types¹², which contain networks of Si and PtSi basis

chains that correspond to that for the polyacetylene basis chains in the hinged polyacetylene phase. As these phases are crowded, especially because of the presence of interstitial atoms (Th and La in the above examples), stretch-densification is unlikely. However, if these basis chains are sufficiently rigid compared to the effective modulus for the hinge deformation mode, both auxetic behaviour and negative thermal expansion would be expected. Wu *et al.*²⁰ have recently synthesized a large organic molecule which is a quantum dot analogue of the hinged phases described here. This molecule provides a soft shear mode due to C—C≡C—C hinges that couple together poly-*m*-phenylacetylene basis-chain segments, which would result in the auxetic property on at least a localized molecular level. Extension of this motif in three dimensions, and using either phenyl substituent groups or intercalants of various sizes, would provide a host of both crowded and uncrowded twisted-chain auxetics. □

Received 13 July; accepted 7 September 1993.

1. Evans, K. E., Nkansa, M. A., Hutchinson, I. J. & Rogers, S. C. *Nature* **353**, 124 (1991).
2. Gibson, L. J., Ashby, M. F., Schajer, G. S. & Robertson, C. I. *Proc. R. Soc. A* **382**, 25–42 (1982).
3. Lakes, R. S. *Adv. Mater.* **5**, 293–296 (1993); *J. Mater. Sci.* **26**, 2287–2292 (1991).
4. Yeganeh-Haeri, A., Weidner, D. J. & Parise, J. B. *Science* **257**, 650–652 (1992).
5. Lakes, R. S. *Science* **235**, 1038–1040 (1987).
6. Guo, Y., Karasawa, N. & Goddard, W. A. *Nature* **351**, 464–467 (1991).
7. Allinger, N. L. & Sprague, J. T. *J. Am. chem. Soc.* **95**, 3893–3900 (1973).
8. Kao, J. & Allinger, N. L. *J. Am. chem. Soc.* **99**, 975–986 (1977).
9. Heravi, M. J., McManus, S. P., Zutauf, S. E. & McDonald, J. K. *Macromolecules* **24**, 1055–1063 (1991).
10. Allen, F. H. *et al.* *J. chem. Soc. Perkins Trans. II* **S1**–S19 (1987).
11. Wegner, G. *Naturforsch.* **24b**, 824–832 (1969).
12. Klepp, K. & Parthé, E. *Acta crystallogr.* **B38**, 1105–1108 (1982).
13. Hoffmann, R., Hughbanks, T., Kertesz, M. & Bird, P. H. *J. Am. chem. Soc.* **105**, 4831–4832 (1983).
14. Liu, A. Y. & Cohen, M. L. *Phys. Rev. B* **43**, 6742–6745 (1992).
15. Brédas, J. L., Chance, R. R., Baughman, R. H. & Silbey, R. *Int. J. Quantum Chem.* **15**, 231–241 (1981).
16. Batchelder, D. N. *Contemp. Phys.* **29**, 3–31 (1988).
17. Koshihara, S., Tokura, Y., Takeda, K. & Koda T. *Phys. Rev. Lett.* **68**, 1148–1151 (1992).
18. Baughman, R. H. *Makromol. Chemie Macromol. Symp.* **51**, 193–215 (1991).
19. Diederich, F. & Rubin, Y. *Angew. Chem. Int. Edn engl.* **31**, 1101–1123 (1992).
20. Wu, Z., Lee, S. & Moore J. S. *J. Am. chem. Soc.* **114**, 8730–8732 (1992).

ACKNOWLEDGEMENTS. We thank H. Eckhardt for important early contributions to this work, C. Cui, R. Lakes and W. B. Hammond for suggestions, and H. L. Baughman for the construction of molecular models. D.S.G. acknowledges financial support from the Brazilian agency FAPESP.

Mechanism of the 1991 eruption of Hekla from continuous borehole strain monitoring

Alan T. Linde*, Kristjan Agustsson†, I. Selwyn Sacks* & Ragnar Stefansson†

* Department of Terrestrial Magnetism, Carnegie Institution of Washington, 5241 Broad Branch Road, NW, Washington DC 20015, USA

† Icelandic Meteorologic Office, Bustadavegur 9, 150 Reykjavik, Iceland

VOLCANOES erupt when the pressure in a magma chamber several kilometres below the edifice overcomes the strength of the intervening rock. Seismic activity may accompany and precede eruptions, allowing (in favourable circumstances) the location and movement of magma to be traced. Ground deformation near volcanoes can provide more direct evidence for magma movement, but continuous monitoring is necessary to ensure that all the essential aspects of an eruption are recorded. Here we report dilatational strain data collected continuously during the January 1991 eruption of Hekla volcano by five borehole strainmeters located 15–45 km from the volcano. The data record the upward propagation of magma, as

well as the deflation of a deep reservoir. In only 30 minutes the magma forced open a conduit to the surface from a depth of ~ 4 km. Although other volcanoes might behave differently, our results suggest the possibility of using continuous deformation measurements to monitor conduit formation at other sites, perhaps providing short-term warnings of impending eruptions.

The Sacks–Evertson strainmeters¹ were installed in southern Iceland in 1979 (Fig. 1) as part of a programme to monitor the transform zone, a region of high seismic potential. Recent tectonic activity includes a magnitude 5.8 earthquake just south of Hekla in 1987 and three small eruptions of Hekla in 1970, 1980 and January 1991. The last eruption was well recorded by all strainmeters, the closest being 14.6 km from the summit. The others are at distances between 37 and 45 km and cover a wide azimuthal range. Previous work^{2,3} suggests that the magma chamber beneath Hekla has a centroid depth of 7–9 km; the data available for those studies allowed modelling of only a spherical pressure source at depth. Our continuous strain data (16-bit resolution, sampling interval 1.4 s) provide the additional capability of determining parameters of the conduit forced open to allow magma to escape to the surface as well as an estimate of the speed of magma ascent. Monitoring of Kilauea, Hawaii⁴ has provided estimates of the speed of horizontal propagation of dikes. Deformation data showed large changes over intervals of days, interpreted as dike emplacement, before a submarine eruption near Ito, Japan^{5,6} but those studies did not give an estimate of propagation speed.

Figure 2 shows data for five days covering the eruption. The overall nature of the strain changes, with most of the deformation occurring during the first several hours of the eruption, is consistent with observations of the rate of lava production⁷. The strong coherence between the distant stations indicates that we should be able to understand the major features of the eruption in terms of rather simple modelling. Because the strain recorded at strainmeter BUR (see Fig. 1) undergoes a reversal in sign ~ 1.5 h after the beginning of the eruption (Figs 2, 3), a model in which the only change is loss of pressure in a deep source is inadequate. Such a process necessarily results in a unidirectional

strain change at any particular site. We therefore use a slightly more complicated model, consisting of deflation of a deep spherical source together with the opening of a conduit (dike) to the surface.

The surface breakout of the eruption is not determined precisely by visual or other direct observation (because of inclement weather conditions); airport radar observations set the time between 17:00 and 17:10 (when the plume was already 11.5 km high) and farmers near the volcano report seeing activity at $\sim 17:05$ (ref. 7; all times are local). Clear strain changes due to magma movement are evident before 16:40. The rising magma forces open a conduit on its way to the surface producing contraction (negative strain) at BUR (deflation of the reservoir results in expansion at BUR). The strain-rate amplitude (Fig. 3b) increases until the time of inferred surface breakout (17:01) and then contraction accumulates at a slower rate. This is because the surface opening releases pressure in the conduit, slowing its opening. Also shown in Fig. 3 are the times of the first-occurring located earthquakes associated with the eruption. These small events (magnitudes < 1) are presumably due to increased stresses caused by initiation of conduit formation at the top of the magma reservoir. The first of these occurred at about the onset time of the strain change at BUR. The focal depths of the first two earthquakes are 5 ± 1 km; because these events are at the edge of the seismic net⁸, the depth estimates were improved by using, in addition to the usual direct-path seismic waves, the travel times of surface reflected waves. Smaller earthquakes may have occurred, but the detection threshold is less than magnitude 0.5 (ref. 8).

We base our calculations on expressions derived for an elastic half space. Okada⁹ has determined surface deformations due to the opening of a rectangular source. The deformation produced by a dike of given geometry and location depends on its volume (length $\times$ width $\times$ opening); that is, volume is a measure of the strength (capacity for producing deformation) of the dike. Deformations due to changes in the magma reservoir are adequately modelled in terms of a pressure source¹⁰, with the strength in this case being the product of the pressure change and source volume. With but five observational sites, only restricted inversions are possible. To ensure robustness, we restrict the range of geometric parameters considered by taking the simplest geometry consistent with local tectonics. We assume that the deep source is located directly beneath the volcano summit and that the conduit to the surface is elongated N60°E, the approximate orientation of the main eruptive fissure at Hekla. It then remains to determine the centroid depths and strengths of the two sources (the deep reservoir and the dike). The eruption, which was essentially complete within two days⁷, may be divided into two stages. During stage 1, both sources contribute to the observed strain changes; the deep reservoir undergoes a loss of magma (pressure decrease) into the conduit which increases in size. At BUR the deep source produces expansion while the dike causes contraction. Clearly the dike continues to increase in strength until at least the time of the minimum at BUR at 18:40 on 17 January. The rapidity of strain reversal at BUR is indicative of little continuing dike growth. We take the time of the BUR minimum to mark the end of growth of the dike. Subsequent deformation (stage 2) is then due solely to continued deflation of the deep reservoir.

To evaluate parameters, we start with stage 2. A spherical pressure source produces surface deformation with circular symmetry and radial variation dependent on the depth. Thus ratios of stage-2 changes at BUR to those at the other sites constrain the allowable depth range for this source, while the magnitudes of strain changes determine the strength. We examined a range of depths, calculating the source strength at each depth which gave the best fit to the data. The data are satisfied (to the extent shown in Fig. 2) by a reservoir with centroid depth 6.5 km and a pressure decrease of 14 MPa for an assumed radius of 2.5 km. A ± 1 km change in source depth results in a 30% misfit for the

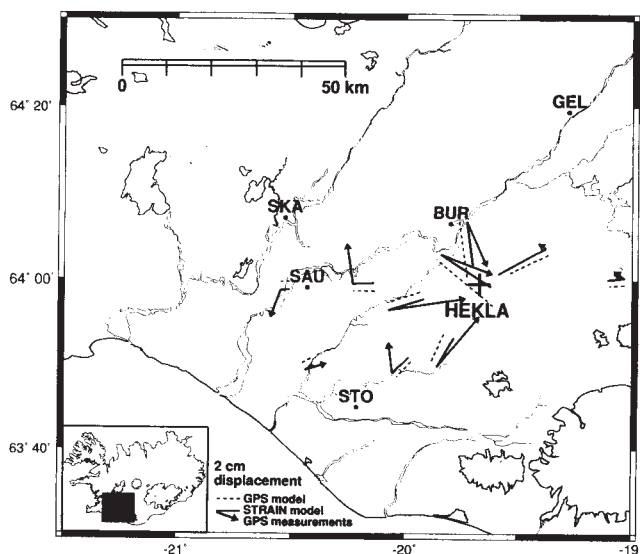
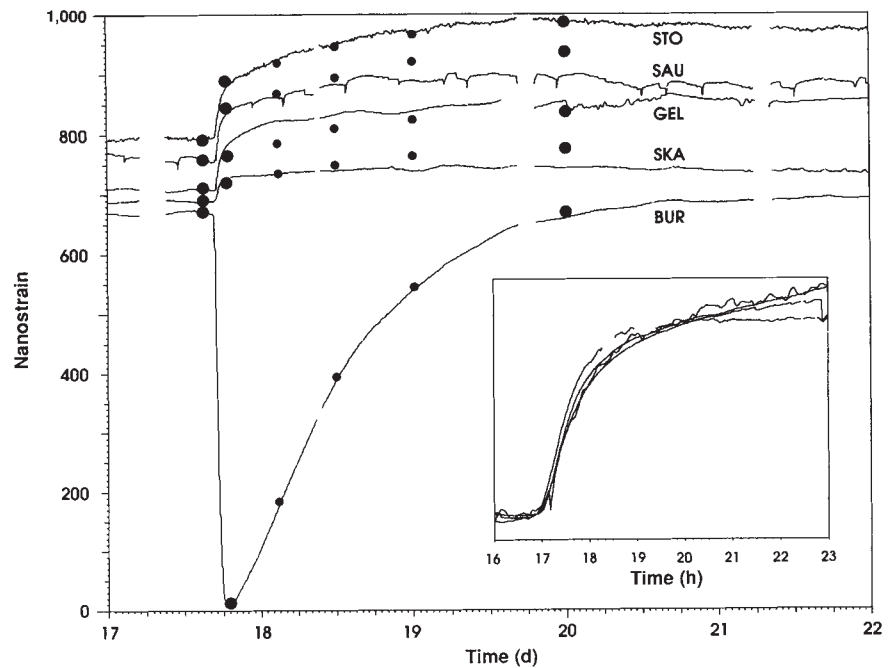


FIG. 1 Map of southwest Iceland showing the location of the borehole strainmeters (three-letter codes) and Hekla volcano (+). Also shown are the sites at which GPS measurements were made, the observed horizontal displacements (arrows) referenced to zero change at Reykjavik, values calculated both from our model (solid lines) and from that of Sigmundsson *et al.*³ (dashed lines), the latter being slightly offset from the observation points for clarity.

FIG. 2 Five days of data from the borehole instruments (day 1=1 January). The original data were low-pass filtered and resampled at 1-min intervals. Expansion is positive. Earth tides and atmospheric-pressure-induced strain changes have been removed. Inset (covering 7 h, and with amplitudes normalized) shows that the more distant sites undergo expansion in a remarkably coherent manner. Large solid circles show the strain changes calculated from modelling the eruption in two stages. Stage 1 (deflation of a deep reservoir together with dike formation) takes place during the interval from initiation of magma movement until the output from strainmeter BUR reaches a minimum. We model stage 2, the following 2 days during which the eruption is essentially complete, solely by continued deflation of the deep source. Smaller circles show the results of calculations for intermediate times during stage 2. These three points correspond to pressure decreases of 30, 60 and 80% of the total change of 14 MPa during stage 2.

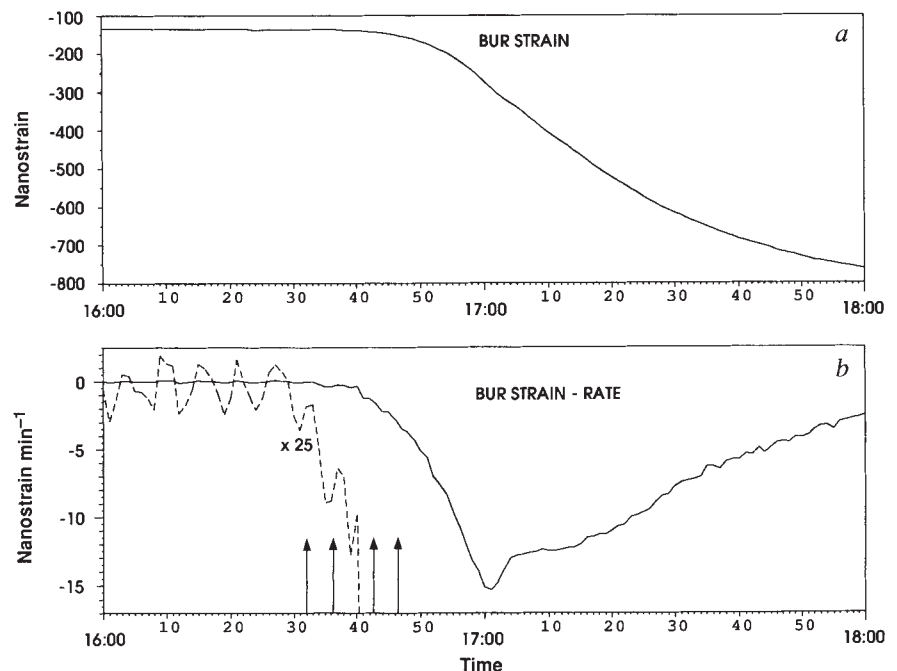


BUR signal. The depths of 5 ± 1 km for the small earthquakes are consistent with the top of the deep reservoir at 4 km. This also sets the depth extent of the dike to 4 km. We use this geometry for stage 1 and vary the strengths of the two sources to fit those strain changes (Fig. 4). We calculate a pressure decrease in the deep reservoir of 14 MPa with a dike 4 km long, 4 km deep and opening 85 cm, consistent with the observed thicknesses of dikes in the area¹¹. A 10% change in dike strength gives a 20% variation in resultant strain change at BUR. The quality of fit to the data makes us confident that the model reflects the essential aspects of the magma movement. Discrepancies between data and calculated values could be reduced either by allowing changes in source horizontal coordinates and in dike strike, or by adding complexity to the model. We have refrained from introducing a multitude of extra parameters.

An additional constraint on the strength of the deep source derives from the erupted volume of lava, estimated⁷ as 0.15 km^3 . We calculate 0.1 km^3 loss in volume of the deep reservoir (volume of the dike is negligible) for a bulk modulus of 17.5 GPa ¹². The density of the lava is $\sim 30\%$ less than that of the magma¹³ due to vesiculation and scoria. As mass is the quantity to be conserved, these two volume estimates are in excellent agreement.

We have also used our model to calculate horizontal displacements at sites monitored with Global Positioning System (GPS) receivers³ (see Fig. 1). Our calculated values match those observed about as well as do those calculated by Sigmundsson *et al.* (ref. 3). GPS surveys were conducted about 18 months before and about a month after the eruption, so the data necessarily include horizontal deformations that occurred before the

FIG. 3 Two hours of strain data (a) and strain-rate data (b) from strainmeter BUR: the x-axis shows local time on 17 January. The dashed strain-rate curve is magnified 25 times. The minimum in the strain-rate gives the inferred time of surface breakout (see text). Also shown by arrows are the times of the first-occurring located earthquakes associated with the eruption; the first of these occurs at about the time strain changes at BUR indicate the beginning of magma movement.



eruption. At some sites, other processes clearly dominate the measurements. Additionally, errors in the displacement data are > 1.5 cm so the discrepancies between data and modelled values are not surprising. The two sets of model values are comparable because horizontal displacements at the GPS sites are largely controlled by the deep source, which is all that was modelled by Sigmundsson *et al.* That our model, determined only on the basis of strain changes during the major volcanic activity, agrees well with the GPS data implies that inflation of the main reservoir during the 18 months before the eruption is small compared with the co-eruption deflation. Such changes are more difficult to detect in our data. Earth noise increases with decreasing spectral frequency and our nearest site is 15 km from the summit resulting in small signals relative to those obtainable closer to the volcano.

The data set discussed here has provided an unprecedented opportunity to determine the mechanism of a volcanic eruption: we are able to monitor the movement of magma from the deep reservoir and determine the source characteristics and changes in them during the eruption. In addition, because the magma source was at depths below ~ 4 –5 km, we know that the ascent velocity of the magma is ~ 10 km h^{-1} . Similar values apply for the eruptions in 1970 (ref. 14) and 1980 (ref. 15) if we assume that the first recorded earthquakes mark the initiation of magma ascent.

Caution should be exercised in generalizing these results because of the different ways in which volcanoes erupt. Magma may make its way to the surface through a more cylindrical

conduit rather than a dike, and there are significant variations in the gas content of magmas. Despite these real differences, we may be optimistic that this type of observation would provide valuable information for other volcanoes as eruptions must be preceded and accompanied by magma forcing a path from a reservoir, at some depth, to the surface. This movement, whether through dikes or otherwise, deforms the surrounding rock and necessarily generates strain changes detectable by near instruments. The time interval during which we might be able to observe deformation before the eruption will depend directly on the depth to the reservoir. $\square$

Received 12 May; accepted 7 September 1993.

1. Sacks, I. S., Suyehiro, S., Evertson, D. W. & Yamagishi, Y. *Pap. Met. Geophys.* **22**, 195–208 (1971).
2. Kjartansson, E. & Grönvold, K. *Nature* **310**, 139–141 (1983).
3. Sigmundsson, F., Einarsson, P. & Björnsson, R. *Geophys. Res. Lett.* **19**, 1483–1486 (1992).
4. Tilling, R. I. & Dvorak, J. J. *Nature* **363**, 125–133 (1993).
5. Shimada, S. *et al. Nature* **343**, 631–633 (1990).
6. Okada, Y. & Yamamoto, E. *J. Phys. Earth* **39**, 177–195 (1991).
7. Gudmundsson, A. *et al. Bull. volcan.* **54**, 238–246 (1992).
8. Stefansson, R. *et al. Bull. seism. Soc. Am.* **83**, 696–716 (1993).
9. Okada, Y. *Bull. seism. Soc. Am.* **75**, 1135–1154 (1985).
10. Yang, X., Davis, P. M. & Dieterich, J. H. *J. geophys. Res.* **93**, 4249–4257 (1988).
11. Gudmundsson, A. *Tectonophysics* **176**, 257–275 (1990).
12. Sato, H. & Manghni, M. H. *Phys. Earth planet. Inter.* **41**, 18–33 (1985).
13. Pálsson, S., Haraldsson, G. I. & Vigfússon, G. V. Report No. OS-84048/VOD-18 B (National Energy Authority, Reykjavik, 1984).
14. Thorarinnsson, S. & Sigvaldason, G. E. *Bull. Volcan.* **36**, 269–288 (1973).
15. Grönvold, K., Larsen, G., Einarsson, P., Thorarinnsson, S. & Saemundsson, K. *Bull. Volcan.* **46**, 349–363 (1983).

ACKNOWLEDGEMENTS. We thank P. Segal for critically reading the manuscript.

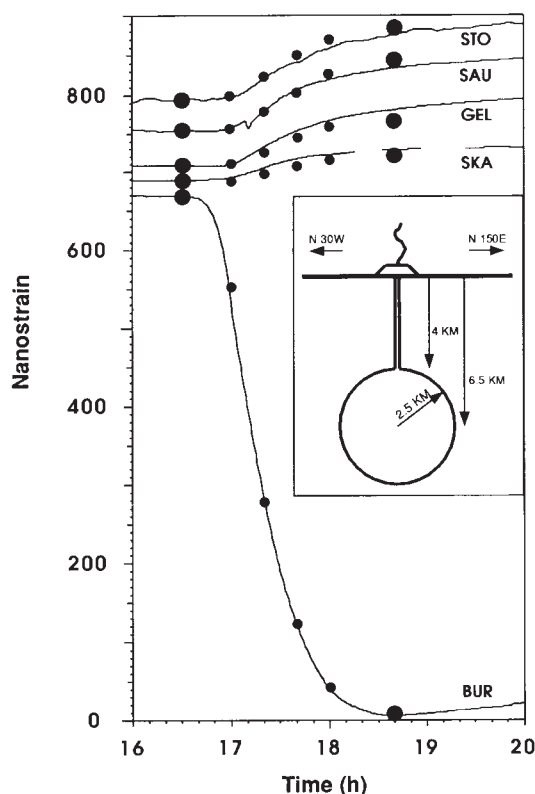


FIG. 4 Four hours of strain data (hour 1 is 01:00 on 17 January). The inset is a sketch of the model used for calculation, with cross-section perpendicular to the strike direction (N60°E) of Hekla and its main fissures. Large circles indicate the pre-eruption levels and the model strain changes calculated at the end of stage 1 when dike formation is complete (85 cm opening, 4 km long, 4 km wide) and the deep reservoir pressure change is 14 MPa. Smaller circles show the results of intermediate calculations. These points correspond to 10, 40, 70 and 90% of dike growth, together with 5, 25, 55 and 85% of stage-1 pressure decrease in the reservoir.

Grain size segregation and stratigraphy in aeolian ripples modelled with a cellular automaton

Robert S. Anderson* & Kirby L. Bunas†

*Department of Earth Sciences, †Computational Math Program, University of California, Santa Cruz, Santa Cruz, California 95064, USA

AEOLIAN ripples are distinguished by coarse-grained crests, fine-grained troughs, and thin veneers of coarse grains on the upwind (stoss) slopes. Migration of these sorted bedforms during periods of net deposition results in a characteristic inversely graded stratigraphy that distinguishes aeolian sandstones from those of fluvial origin¹. Here we investigate the formation and evolution of aeolian ripples using a cellular automaton model of a sandbed which incorporates two grain sizes and is subject to the impacts of episodically hopping (saltating) grains. Using this model, we identify the physical processes responsible for both the spatial sorting and the stratigraphic signature of the ripples. High-energy saltating grains are found to eject small grains preferentially from the impact site, leading to a coarsening of the heavily bombarded stoss slopes. Coarse grains do not in general leap far enough to escape the shadow zones on the downwind (lee) slopes, and therefore tend to accumulate and cycle around the ripple crests. Small grains, on the other hand, are ejected at higher velocities, enabling them to hop further. Incorporating net deposition into the model results in a stratigraphy closely resembling thin planar laminae ('pinstriping') found in aeolian sandstones: the inverse grading of grain sizes is a direct consequence of their different hop lengths.

Aeolian ripples represent one of the more spectacular and ubiquitous self-organized systems in terms of both topography and texture². These smallest of bedforms arise when sand is transported in energetic hops that periodically strike the bed in

the process called saltation³. Flat beds are unstable with respect to slight topographic perturbations⁴ that grow through time into mature forms that migrate downwind. Mature ripples are asymmetrical (Fig. 1), with convex upwind (stoss) slopes of 10° and concave lee slopes of angles up to the angle of repose (30°) but typically of 20° (refs 5, 6). Grain sizes are significantly segregated, with coarse grains on the ripple crests, an effect for which no satisfactory explanation exists. Ripple stratigraphy in the rock record is characterized by abrupt, slightly upwind-dipping planar bounding surfaces between which the laminae are inversely graded (coarse over fine). This stratigraphic signature has been attributed to the translation of sorted ripples during periods of net deposition¹.

Previous numerical models of ripple formation⁷⁻⁹ have focused on ripple wavelength and shape evolution, and have treated only single-grain-sized beds. An important element in the ripple process is the pattern of saltation impact intensity on the evolving form, with highest intensities on the stoss slope. Near the top of the lee face on mature ripples, where the topographic slope is greater than the 10° saltation impact angle, the surface is entirely shadowed from impacts. These models have also shown that ripples grow in wavelength through time by means of the merger of ripples moving at different speeds⁷⁻⁹.

The only previous attempt to model the development of ripple stratigraphy involved a single-grain-size cellular automaton model in which surface grains were periodically 'stained'⁸, creating an effect akin to that of periodically sprinkling coloured grains on the rippled surface.¹ Instead, we model the ripple process using a two-grain-size cellular automaton in which each grain occupies a single cell, and in which the grain impact, the ejected grain trajectory, and the probability of deposition of the incoming grain are described by realistic rules detailed below. By explicitly incorporating two grain sizes, we directly assess the mechanisms responsible for ripple sorting and ripple stratigraphy.

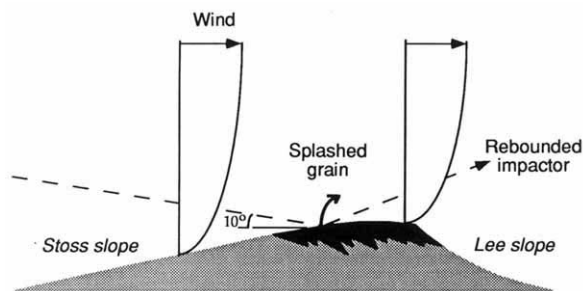


FIG. 1 Schematic cross-section of a mature natural aeolian ripple. The asymmetrical profile, with low-angle stoss and high-angle lee slopes^{5,6}, and the coarsened crest (shown in black) are typical. One low-angle saltation impact is depicted, with high probability rebound of the impactor and near-simultaneous 'splash' of low-energy grains from a region near the impact point. Low-angle impacts allow significant modulation of the impact intensity over finite amplitude ripples: stoss slopes are bombarded heavily, lee slopes lightly, with impact shadow zones immediately downwind of ripple crests. The wind velocity profile over two points on the bed is shown. The flow profile over an initially flat bed is logarithmic: $U = (u^*/k) \ln(z/z_0)$, where u^* is the shear velocity, k is von Karman's constant, z is the height above the bed and z_0 is the roughness height. Bed topography causes pinch-and-stretch of this flow. In our model, this is controlled by setting a 'ceiling height' defined as the altitude above the mean bed below which significant divergence and convergence of flow lines occurs, and above which the flow field is horizontally uniform. We set the ceiling height at one mean ripple wavelength, a choice justified by calculations of the potential flow field over a sinusoidal boundary. While this flow model only crudely approximates real flows that involve a complex topographically-induced structure of both mean and fluctuation (turbulence) velocity components, it captures the essence of the flow pattern relevant to the trajectories of near-bed particles.

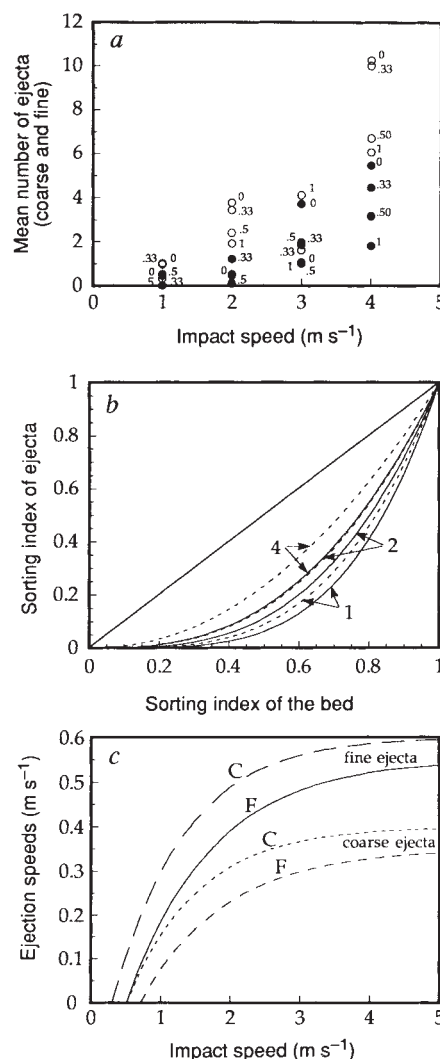
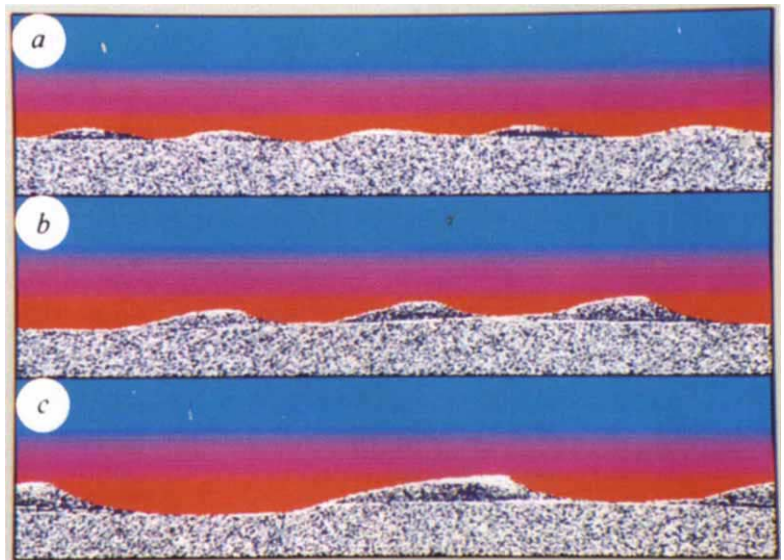


FIG. 2 Statistical description of the multiple-grain-size impact derived from a particle dynamics code¹⁰. Statistics and curve fits to data from impacts into all-fine and all-coarse beds are from ref. 10; remaining statistics on beds with coarse/fine ratios of 1/2 and 1/1 are from unpublished data (P. K. Haff and R.S.A.). a, Numbers of ejected particles (ejecta) per impact as a function of impact speed. Numbers beside symbols correspond to sorting index (fraction of coarse grains) of the bed. ○, coarse impactor; ●, fine impactor. As in the single-grain-size cases, higher energy (large grains, high speeds) impacts result in larger numbers of ejecta¹⁰, with the number of ejecta increasing roughly linearly with impact speed over this range of typical saltation impact speeds. Curve fits to these data were used in the simulations. b, Curves fitted to the sorting index (or ratio) δ in the ejected population (δ = number of coarse grains ejected/total number ejected) as a function of the sorting index S of the bulk bed (S = coarse/total). Numbers correspond to impact speed in m s⁻¹. Dashed lines indicate coarse impactor, solid lines fine. All impacts preferentially result in a lower coarse/fine ratio in the ejected population than in the local bed; all curves fall below the 1/1 line. The resulting fractionation of the bed is greatest for low-energy impacts (small grains, low speeds) that cannot eject large grains, and is smallest for high-energy impacts (large grains, high speeds) that indiscriminately blast away the local bed. c, Curves fitted to results from impact code¹⁰ for ejection speed against impact speed for mixed-grain-size beds. Letters C (coarse) and F (fine) correspond to the impactor size. Results show the significantly higher ejection speeds of the fine grains, and slightly higher mean ejection speeds resulting from coarse (compared to fine) impactors. The ejection speed seems to reach an asymptotic value (fine grains, 0.6 m s⁻¹; coarse grains, 0.35 m s⁻¹) at high impact speeds; further increases in impact energy result instead in increasing numbers of ejecta (see Fig. 2a). The roughly 5/3 ratio of fine/coarse ejection speeds results in 2.5/1 ratio (fine/coarse) of expected trajectory heights, given that the ejection angles are approximately independent of grain size.

FIG. 3 History of ripple development from initial flat bed to mature, well sorted ripples. Fine grains (0.23 mm) are dark blue; coarse grains (0.32 mm) are white. Wind blows left to right with $u^* = 0.5 \text{ m s}^{-1}$; ceiling height (see Fig. 1 legend) = mean ripple wavelength. The bed is initially 1,250 grains wide and 100 grains deep. The grains are randomly assigned sizes, with a fine/coarse ratio of 2/1. The grains that strike the bed have an identical size distribution; impact angles α are randomly selected from a gaussian distribution with $\bar{\alpha} = 11^\circ$, $\sigma = 2^\circ$ and speeds v_{im} from values above u^* within $p(v_{im}) = (1/u^*)\exp(-v_{im}/u^*)$. The splash rule includes the total number, size distribution and speeds of ejecta from analytical fits to the data shown in Fig. 2. Neither aerodynamic entrainment nor an angle of repose condition is incorporated; splashed grains stick where they land. Ripples are depicted at 5 million (a), 10 million (b) and 20 million (c) impacts. Mean wavelength grows rapidly at first, more slowly later. Significant coarsening of crests occurs once the ripples have become established. Stoss slopes all display a thin but ubiquitous layer of coarse grains, reflecting impact fractionation. Impact rates at $u^* = 0.5 \text{ s}^{-1}$ (refs 13, 15) imply that the evolution shown here occurs in several minutes of real time, in accord with wind-tunnel observations⁷. While the temporal evolution of ripple form therefore mimics well the natural growth of ripples, the evolution of ripple sorting has not been documented in the real world.



Statistics derived from modelling the dynamics of individual granular impacts (ref. 10, and P. K. Haff and R. S. A., unpublished data) were first used to construct a set of rules quantifying the process of saltation impact into a bed of mixed grain sizes (Fig. 2). In the mixed-grain-size case, as with a single grain size^{10,11}, energetic grains tend to rebound at $\sim 50\%$ of their incoming speed. A fraction of the impact energy results in grain ejection near the impact site, with velocities $\sim 10\%$ of the impact speed. The total number of ejected grains (Fig. 2a), the ratio coarse/fine in the ejecta (Fig. 2b) and their velocities upon ejection (Fig. 2c) depend on the size and speed of the impactor and upon the sorting of the local bed. Fine grains are preferentially ejected. Statistically, the coarse/fine ratio in the ejected grain population is always less than or equal to that in the local bed (Fig. 2b). The difference is greatest at low impact speeds and with small impactors, when the impactor imparts too little energy to move a large grain, and is least at high impact speeds with large impactors, when enough energy is available to eject the entire local population indiscriminantly. As most saltation impacts are below 4 m s^{-1} (refs 12, 13) this impact process consistently results in 'impact fractionation' of the bed, leaving the local bed coarser. In addition, smaller grains are ejected with higher speeds¹⁰ (Fig. 2b). Ejection angles are roughly independent of grain size.

Our model bed is composed of two grain sizes, coarse (0.32 mm) and fine (0.23 mm) in a 1/2 ratio, is 1,250 grains wide (0.325 m) and initially 100 grains deep. Saltating grains are fired

at the bed one by one. The grain size, the speed and the angle of impact are chosen stochastically from gaussian distributions, all tied to the wind velocity using rules developed from earlier work on saltation^{12,13}. The granular splash at the point of impact is described by the above rules (Fig. 2), resulting in initial conditions for the ejected grain trajectories. The subsequent trajectory is determined by the grain mass and the wind velocity field. In general, the higher speeds of the small grains result in significantly longer trajectories. As hop height is roughly proportional to the square of the initial vertical velocity¹⁴, small grains access the higher wind velocities at greater heights above the bed. Lighter grains also experience greater downwind acceleration. Both effects increase hop lengths of small grains.

As in earlier single-grain-size ripple models^{7,9}, small-wave-length mottles of the bed grow and coalesce into mature ripples (Fig. 3). Rate of growth of the mean wavelength diminishes through time. In contrast to earlier models, however, the forms of the mature ripples are asymmetric, as in natural ripples (Fig. 3). Neither stoss nor lee slopes are straight; rather, the ripples display the convex stoss slopes and concave lee slopes characteristic of natural ripples^{5,6}. Ripple shape is sensitive to the details of the splash trajectory calculations. The (natural) asymmetric forms arise only when grain trajectories are explicitly calculated (rather than being estimated for calculational ease) and when the wind field is considerably compressed over crests and stretched over troughs (by setting a 'ceiling height' at roughly one wavelength above the mean bed; see Fig. 1). Simulations in

FIG. 4 Cellular automaton results mimicking "subcritically climbing translant ripple strata"¹ after 2×10^7 impacts. Grain colouring as in Fig. 3; wind blows left to right. Only 0.6% of the grains that strike the surface are deposited, the deposition probability being inversely proportional to speed of impact. As the slower coarse grains are preferentially captured, the ratio of coarse/fine in the impacting population is set to 1/13 in order to maintain the bed ratio of 1/2. The lower part of the figure shows the original randomly sorted bed. Above this, packages of inversely graded sand bounded by relatively straight upwind-dipping surfaces fit well the description of "subcritically climbing translant strata"¹; these strata contain faint concave-up stratification akin to "ripple fore-set cross laminae" that are found, similarly poorly developed, in aeolian sandstones¹.



which this effect was artificially turned off (all flow lines parallel to the bed) result in reduced asymmetry, whereas simple estimations of trajectory lengths result in steep and almost triangular bedforms similar to those reported from earlier numerical models⁸. We emphasize that no grains are aerodynamically entrained from the bed; grains are not ripped from the ripple summits, but are rather 'slingshot' past them by the pinch and swell of the flow field.

The mature ripples display strong segregation of grains (Fig. 3). Coarsened crests occur only after the ripples have reached significant heights, when lee slopes are strongly shadowed from impacts. We also find that stoss slopes are always coarsened (Fig. 3). Although not often reported, presumably because this thin film of coarse grains is difficult to sample, this effect is clearly observable in natural ripples. We argue that this stoss coarsening is a straightforward manifestation of impact fractionation. When both coarse and fine grains are forced to be ejected with identical distributions of speeds, and so hop roughly equal distances, the coarse crests vanish. The differential hop-length effect (that we consider is the cause of the segregation of coarse grains into ripple crests) may be explained as follows. Coarse grains travel in very small hops of the order of one to a few grain diameters. On leaping over the crest of a mature ripple with a well developed impact shadow zone, coarse grains can not leap far enough down the lee side to escape this shadow, and the slope is not steep enough to allow grains to roll to the base of the lee. Coarse grains are then buried by subsequent deposition on the lee face, exhumed high on the stoss face on downwind translation of the ripple, and once again skitter up to the crest to repeat the process. The coarser the grain, the greater this effect.

Deposition of some fraction of the impacting grains that strike the surface is required to cause net accretion of the bed and hence create ripple stratigraphy. Our deposition rule forces preferential capture of the lowest-speed and highest-angle impactors, which will tend to be the coarser grains. To maintain a bed with a coarse/fine ratio of 1/2 therefore requires a saltating population heavily skewed toward the fine grains, as is found in field experiments¹⁵. The resulting stratigraphy closely mimics 'subcritically climbing' ripple stratigraphy¹ (Fig. 4). It is characterized by thin planar packages of inversely graded sand between bounding surfaces that dip slightly upwind. Wispy downwind-dipping cross-bedding between bounding surfaces is associated with slight random variations in coarse and fine grain populations landing on the lee faces, and represent time-lines in the deposit. On the other hand, the bounding surfaces are not time lines; they record the downwind and slightly upward translation of that portion of the stoss face where the slope equals the angle of climb of the strata. Because stoss slopes are always coarsened by impacts, the sand immediately below the bounding surfaces is always slightly coarser than the underlying deposit.

We conclude that the essence of aeolian ripple sorting and stratigraphy lies in the multiple-grain-size splash, which is the cause of both the impact fractionation and the hop-length differentiation processes. This long-standing geomorphic and stratigraphic puzzle was solved here by linking short timescale simulations of the impact process¹⁰ with a longer timescale model that allows efficient exploration of the variety of available mechanisms through real-time visualization of the evolving spatial patterns. □

Received 21 June; accepted 27 August 1993.

1. Hunter, R. E. *Sedimentology* **24**, 361–387 (1977).
2. Hallet, B. *Earth Sci. Rev.* **29**, 57–76 (1990).
3. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Methuen, London, 1941).
4. Anderson, R. S. *Sedimentology* **34**, 943–956 (1987).
5. Sharp, R. P. *J. Geol.* **71**, 617–636 (1963).
6. Werner, B. T., Haff, P. K., Livi, R. P. & Anderson, R. S. *Geology* **14**, 743–745 (1985).
7. Anderson, R. S. *Earth Sci. Rev.* **29**, 77–96 (1990).
8. Forrest, S. B. & Haff, P. K. *Science* **255**, 1240–1243 (1992).
9. Werner, B. T. thesis, California Inst. of Technol. (1987).
10. Haff, P. K. & Anderson, R. S. *Sedimentology* **40**, 175–198 (1993).

11. Werner, B. T. *J. Geol.* **98**, 1–17 (1990).
12. Anderson, R. S. & Haff, P. K. *Science* **241**, 820–823 (1988).
13. Anderson, R. S. & Haff, P. K. *Acta mech. (Suppl.)* **1**, 21–50 (1991).
14. Anderson, R. S. & Hallet, B. *Geol. Soc. Am. Bull.* **97**, 523–535 (1985).
15. Anderson, R. S. & McDonald, R. R. *Sedimentology* (submitted).

ACKNOWLEDGEMENTS. This research was supported by the US National Science Foundation. We thank the Computational Math Program of UCSC for supplying the computer facilities. We also thank P. K. Haff for discussions of the granular splash process and R. R. McDonald for help with the trajectory code.

Hyperthermophilic archaea are thriving in deep North Sea and Alaskan oil reservoirs

K. O. Stetter*, R. Huber*, E. Blöchl*, M. Kurr*,
R. D. Eden†, M. Fielder‡, H. Cash§ & I. Vance§

*Lehrstuhl für Mikrobiologie und Archaeenzentrum, Universität Regensburg, Universitätsstrasse 31, 93053 Regensburg, Germany

†CAPCIS Ltd, UMIST, Bainbridge House, Granby Row, Manchester M1 2PW, UK

‡BP Exploration Operating Company Ltd, Farburn Industrial Estate, Dyce, Aberdeen AB2 0PB, UK

§BP Research, Warrensville Research and Environmental Science Center, 4400 Warrensville Center Road, Cleveland, Ohio 44128, USA

HOT springs and hydrothermal vents harbour hyperthermophilic archaea and bacteria with the highest growth temperatures known^{1–6}. Here we report the discovery of high concentrations of hyperthermophiles in the production fluids from four oil reservoirs about 3,000 metres below the bed of the North Sea and below the permafrost surface of the North Slope of Alaska. Enrichment cultures of sulphidogens grew at 85 °C and 102 °C, which are similar to *in situ* reservoir temperatures^{7,8}. Some species were identical to those from submarine hot vents and may have entered the reservoirs in injected sea water. Several enrichments grew anaerobically in sterilized artificial sea water with crude oil as the single carbon and energy source. These hyperthermophiles may be part of novel high-temperature communities and could be responsible for *in situ* bioconversions of crude oil fractions at temperatures previously considered too extreme for biochemical reactions^{4,7,9,10}.

To investigate the presence of hyperthermophilic microorganisms in deep oil-bearing strata, well-head samples upstream of topsides processing facilities from newly emergent reservoir fluids were initially taken from the Thistle offshore oil production platform located in the East Shetland Basin of the North Sea. Later, similar samples were taken from 33 wells on the Prudhoe Bay, Endicott and Kuparuk oil fields on the North Slope of Alaska, USA. These four reservoirs are all flooded by pressurized sea water as part of the planned secondary oil recovery process. The maximum temperatures encountered in them were 105 °C, 110 °C, 102 °C and 71 °C, respectively. At the well-heads at the time of sampling, depending on the flow rate of the individual wells (up to 3.5 m s⁻¹), the emergent production fluids had cooled to between 20 °C and 87 °C. All four reservoirs have a history of H₂S production.

Hyperthermophiles were detected by inoculating sterile, anaerobic culture media with the water phases of the samples and incubating them for 15 to 70 hours at 85 and 102 °C (Table 1). Under the microscope, the positive enrichment cultures from the Thistle samples showed coccoid, irregular cells about 0.5 to 2 µm in diameter. Within cultures enriched at 85 °C in the absence of elemental sulphur (S⁰), a multitude of cells showed a bluish-green fluorescence under ultraviolet light at 420 nm. In addition, these cultures formed large quantities of H₂S, indicating that the fluorescent cocci were *Archaeoglobus*-like sulphate-reducers⁴.

Disc-shaped cells, about 2 μm in diameter, grew up in enrichments incubated at 85 °C with crude oil as the only substrate. The cells were successfully transferred into eight sequential sub-cultures in fresh medium and grew on the hydrophobic fraction of crude oil (Table 1). In the presence of S^0 and yeast extract at 102 °C and 85 °C, non-fluorescent cocci were enriched that formed H_2S by S^0 reduction.

From the Alaskan reservoirs, enrichments at 85 °C from 13 well-head samples contained *Archaeoglobus*-like fluorescent sulphate-reducing cocci. Sixteen different well-head samples yielded cultures of non-fluorescent sulphur-reducing cocci at 85 °C. In addition, enrichment cultures from 7 well-heads contained rod-shaped (about 0.6 μm in diameter and 4 to 8 μm in length) *Thermotoga*-like (eu)bacteria which grew in a mixture with coccoid archaea¹¹. Hyperthermophiles were not recovered from natural Beaufort Sea water, the source water for the Alaskan secondary recovery. After incubation at 85 °C, however, coccoid sulphate- and sulphur-reducers were successfully enriched from samples of concentrated seawater solids (1 m³ of Beaufort Sea water collected on 0.16- μm filter media).

To establish phylogenetic relationships of the hyperthermophiles from the Thistle reservoir, enrichment cultures were probed using DNA/DNA hybridization with hyperthermophiles known to be present in submarine hot vents^{5,12} (Table 1). The presence of the same species in these different environments was indicated by the high degree of hybridization (ref. 5, and data not shown) of various cultures present in the Thistle fluids with the DNA from the species *Archaeoglobus fulgidus*, *A. profundus*, *Thermococcus celer* and *T. litoralis* (Table 1). In addition, a novel species of *Archaeoglobus* ('*A. lithotrophicus*') was obtained from several production wells on the Thistle platform. Under the electron microscope, cells frequently showed wing-like ultraflat protuberances which increase the surface area to volume ratio and possibly improve gas exchange of this hydrogen-utilizing organism (Fig. 1). Enrichments of heterotrophs at 102 °C gave a weak but significant signal with the DNA of *Pyrococcus furiosus*, indicating that there is a novel species of *Pyrococcus* which grows at temperatures between 60 and 103 °C and was present in many well-head samples (Tables 1, 2). The original population densities of hyperthermophiles in the emergent reservoir fluids were in the range 10^4 to 10^7 l⁻¹, indicating severe contamination with archaeal sulphidogens of the genera *Archaeoglobus*, *Thermococcus* and *Pyrococcus* (Table 1). These values translate into the transport of significant quantities of hyperthermophiles to the surface, and in the case of Thistle, subsequent marine discharge in the range 3 to 16 kg per day for each candidate reservoir. The pressure tolerance of isolate *A. fulgidus* TF2 (from production well A11 on the Thistle platform) was investigated by incubating replicate cultures for one week at 82 and 90 °C at 300 and 420

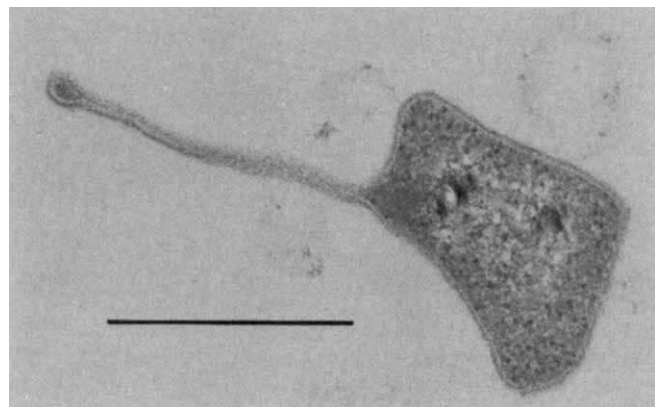


FIG. 1 Ultra-thin section of a budding cell of '*A. lithotrophicus*' TF2 (isolated from production well A11 on the Thistle production platform). Scale bar, 1 μm .

TABLE 1 Hyperthermophiles from Thistle water samples

Well number	Enrichment cultures at 85 °C	102 °C*	Original cell concentrations (viable cells l ⁻¹)	Identified species
A01	+	+	ND	Al;Ti;Ps;Ec
A02	+	+	A 10 ⁵ ; T 10 ⁴ ; P 10 ⁶	Af;Al;Ti;Tc; Ps
A05	+	+	A 10 ⁶ ; T 10 ⁴ ; P 10 ⁴	Af;Al;Ti;Ps; Ec
A07	+	+	A 10 ⁴ ; T 10 ⁴ ; P 10 ⁴	Al;Ti;Tc;Ps
A08	+	+	A 10 ⁷ ; T 10 ⁵ ; P 10 ⁴	Af;Ap;Al;Ti; Ps
A10	+	+	A 10 ⁶ ; T 10 ⁴ ; P 10 ⁴	Af;Ap;Al;Tf; Tc;Ps
A11	+	+	ND	Af;Ap;Al;Ti; Ps;Ec
A19	+	—	ND	Af
A25	+	+	A 10 ⁵ ; T 10 ⁶ ; P 10 ⁴	Af;Al;Ti;Tc; Ps
A27	+	+	ND	Af;Al;Ti;Ps
A31	+	+	A 10 ⁵ ; T 10 ⁴ ; P 10 ⁶	Af;Al;Ti;Tc; Ps;Ec
A32	+	—	ND	Al
A34	+	+	ND	Al;Ti;Ps;Ec
A41	+	+	ND	Af;Al;Ti;Ps; Ec
A44	+	+	ND	Af;Ti;Ps

Samples of oil-water-gas emulsions were taken from the production stream through a stainless steel tube inserted into the production line at the well-head upstream of process facilities. After flushing the tube for 10 min, anaerobic samples were directly filled into sterile 100 ml glass bottles which were then tightly stoppered³. After 1 to 10 min, emulsions separated into 3 distinct phases. The culture media consisted of artificial sea water with various amendments and the cultures were grown anaerobically^{5,22}. Crude oil was washed 5 times, each with 3 vol of water. Phylogenetic relationships were investigated by dot-blot DNA/DNA hybridization of enrichment cultures with that of type species of hyperthermophiles⁵. The original cell concentrations were estimated by inoculating enrichment media with serial dilutions of the water phase from the original samples and identifying the species as described⁵.

*Only *Pyrococcus* detectable; A, *Archaeoglobus*; T, *Thermococcus*; P, *Pyrococcus*; Af, *A. fulgidus*; Ap, *A. profundus*; Al, '*A. lithotrophicus* sp. nov.'; Ti, *T. litoralis*; Tc, *T. celer*; Ps, '*P. sp. lithotrophicus* sp. nov.'; Ec, positive enrichment culture on crude oil at 85 °C (disc-shaped cells); ND, not determined; +, positive; —, no growth.

atmospheres^{4,13,14}. Increased cell concentrations demonstrate that this isolate can grow under the extreme conditions found in the seawater-supported Thistle reservoir at 85 °C and 350 atmospheres⁷.

Our findings indicate that deep oil reservoirs are a novel, non-volcanic, high-temperature environment for hyperthermophilic archaea. The minimum growth temperatures of the oilfield hyperthermophiles (Table 2) suggest that they are unlikely to be active in surface soils, oil and ground water flows at ambient temperatures^{15,16}. Owing to their high growth temperatures, pressure tolerance and highly anaerobic mode of life, the reservoir hyperthermophiles seem to be well adapted to a subterranean environment of geothermally heated porous rock saturated with petroleum fluids and water. Some of the substrates such as acetate and CO_2 , which are used by reservoir organisms, are known to be present *in situ*¹⁷. The predominant mode of nutrition in reservoirs, however, is at present unknown. Molecular hydrogen could be formed *in situ* abiotically or microbiologically^{11,18,19}. The novel hyperthermophiles that grow anaerobically on a component of crude oil may be a critical part of a complex reservoir food net. Recently, (eu)bacterial mesophiles that oxidized alkanes up to C_{20} anaerobically have been isolated²⁰. All of the hyperthermophilic reservoir isolates

TABLE 2 Temperature range of growth and laboratory substrates of the isolates from the Thistle reservoir

Species	Growth temperature (°C)			Laboratory substrates (positive requirement)	Electron acceptor
	min.	opt.	max.		
<i>Archaeoglobus fulgidus</i>	60	83	94	lactate; cell extracts	SO ₄ ²⁻
<i>Archaeoglobus profundus</i>	65	82	90	H ₂ /CO ₂ + acetate; H ₂ /CO ₂ + cell extracts	SO ₄ ²⁻
' <i>Archaeoglobus lithotrophicus</i> sp. nov.'	63	80	89	H ₂ /CO ₂	SO ₄ ²⁻
<i>Thermococcus litoralis</i>	50	88	98	cell extracts; peptone	S ⁰
<i>Thermococcus celer</i>	75	87	93	cell extracts; peptone	S ⁰
' <i>Pyrococcus</i> sp. nov.'	60	92	103	cell extracts, peptone	S ⁰

Pure cultures of the strains were obtained by plating on to culture medium solidified by 0.6 % gelrite (Kelco, San Diego, USA). 10 ml cultures were grown during shaking at 200 r.p.m. in stoppered 120 ml serum bottles⁵. Organic substrates were added at 0.05 % (w/v). Cell extracts consisted of yeast extract (Difco), meat extract (Difco) and *Methanothermobacter fervidus* extract. Growth temperatures were determined as described⁶. Alternative positive substrates are given after the semicolons.

are able to reduce sulphur compounds which may be provided *in situ* by reservoir minerals, crude oil and injected sea water (Table 2). Because these hyperthermophiles produce H₂S they probably contribute to the biological souring of deep, hot reservoirs.

The origin of the reservoir hyperthermophiles is at present unclear. Isolated communities that have survived since the formation of these reservoirs seem unlikely as (1) all of the reservoirs studied had been flooded with sea water; (2) hyperthermophiles are known to survive in cold sea water⁵; and (3) several of the reservoir species are identical to those found in submarine hot vents. Furthermore, preliminary results suggest the presence of low population densities of viable hyperthermophiles in sea water. Although it seems likely that the hyperthermophiles have entered the reservoirs in the injected sea water, it is possible that natural routes such as faults and oil seeps could also be involved. The scale of oil-field operations makes the inoculation of reservoirs highly probable, even if sea water contains a low concentration of viable hyperthermophiles. The reservoirs investigated, for example, are flooded with between 40,000 m³ and 127,000 m³ of sea water per day. Before injection, sea water is treated with biocidal chemicals but such treatments are not 100 % effective, thus allowing the passage of some hyperthermophiles (and other bacteria) into the reservoir. The explanation of our findings does not depend on the existence of a hot, subterranean biosphere as the origin of the reservoir hyperthermophiles²¹. As hyperthermophiles were recovered from the production fluids of four separate oil fields they are probably widely distributed in hot, seawater-flooded petroleum reservoirs. Further investigations are required to explore the distribution and unique metabolic properties of the archaeal communities from the deep reservoirs. □

13. Morita, R. Y. in *Meth. Microbiol.* **2**, 243–257 (1970).
14. Stott, J. F. D. & Herbert, B. N. *J. appl. Bact.* **60**, 57–66 (1986).
15. Wiegand, J. in *Thermophilic Bacteria* (ed. Kristjansson, J. K.) 105–184 (CRC, Boca Raton, Ann Arbor, and London, 1992).
16. Olson, G. J., Dockins, W. S., McFeters, G. A. & Iverson, W. P. *Geomicrobiol. J.* **2**, 327–340 (1981).
17. Carothers, W. W. & Kharaka, Y. K. *Am. Ass. Petrol. Geol. Bull.* **62**, 2441–2453 (1978).
18. Drobner, E., Huber, H., Wächtershäuser, G., Rose, D. & Stetter, K. O. *Nature* **346**, 742–744 (1990).
19. Fiala, G. & Stetter, K. O. *Arch. Mikrobiol.* **145**, 56–61 (1986).
20. Aeckerberg, F., Bak, F. & Widdel, F. *Arch. Mikrobiol.* **156**, 5–14 (1991).
21. Gold, T. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6045–6049 (1992).
22. Stetter, K. O., König, H. & Stackebrandt, E. *Syst. appl. Microbiol.* **4**, 535–551 (1983).

ACKNOWLEDGEMENTS. We thank R. Rachel for electron microscopy, M. Littlewood, C. Scheimer, H. Segerer, J. Schipka and J. Stott for technical assistance, BP Exploration (Alaska) Inc., Arco Alaska Inc. and the BP Exploration Company and their partners in the Thistle Field for access to their production facilities. This work was supported by the Deutsche Forschungsgemeinschaft (Leibniz-Preis), BP Research, the Fonds der Chemischen Industrie and the UK Multiclient Oilfield Reservoir Souring Programme (CAPCIS Ltd).

Walrus-like feeding adaptation in a new cetacean from the Pliocene of Peru

Christian de Muizon

Institut Français d'Études Andines, URA 12 CNRS, Casilla 18-1217, Lima 18, Peru

THE recent discovery in the Southern Hemisphere (15.5 °S) of a walrus-like skull of a toothed whale (odontocete) from the Pisco Formation of southern Peru presents a startling example of convergence and specialization unprecedented among cetaceans. In contrast to other toothed whales, *Odobenocetops peruvianus* has no elongated rostrum but has large, ventrally directed premaxillary alveolar processes housing asymmetrical tusks. The dorsally facing orbits indicate the possibility of dorsal binocular vision which could compensate for the absence of the melon, a rostral organ involved in echolocation. Strong muscle scars on the anterior edge of the premaxillae suggest the presence of a powerful upper lip. It is this feature, together with the morphology of the deep vaulted palate devoid of maxillary teeth, that enables us to hypothesize a convergent feeding adaptation with the walrus which feeds mainly on thin-shelled bivalve molluscs, sucking out the foot and/or siphon and ejecting the shell. The structure of the face and basicranium indicates it was a delphinoid cetacean, probably related to the living beluga and narwhal (Monodontidae).

Received 24 March; accepted 16 August 1993.

1. Woese, C. R., Kandler, O. & Wheelis, M. L. *Proc. natn. Acad. Sci. U.S.A.* **87**, 4576–4579 (1990).
2. Stetter, K. O., Fiala, G., Huber, G., Huber, R. & Segerer, A. *FEMS Microbiol. Rev.* **75**, 117–124 (1990).
3. Stetter, K. O. *Nature* **300**, 258–260 (1982).
4. Stetter, K. O., Lauerer, G., Thomm, M. & Neuner, A. *Science* **236**, 822–824 (1987).
5. Huber, R., Stoffers, P., Cheminee, J. L., Richnow, H. H. & Stetter, K. O. *Nature* **345**, 179–181 (1990).
6. Huber, R. et al. *Syst. appl. Microbiol.* **15**, 340–351 (1992).
7. Eden, R. D., Laycock, P. J. & Fielder, M. *Oilfield Reservoir Souring* (UK Health and Safety Executive OTH 92385, HMSO, 1993).
8. Wadman, D. H., Lamprecht, D. E. & Mrosovsky, I. *J. Petrol. Technol.* **31**, 933–940 (1979).
9. Marsland, S. D., Dawe, R. A. & Kelsall, G. H. *Trans. I chem. Engrs* **68**, A357–364 (1990).
10. Lighthelm, D. J., de Boer, R. B., Brint, J. F. & Schulte, W. M. *Technical Paper SPE 23141, Society of Petroleum Engineers, Offshore Europe Conf. Sept 3–6, Aberdeen*, 369–378 (1991).
11. Huber, R. et al. *Arch. Mikrobiol.* **144**, 324–333 (1986).
12. Joergensen, B. B., Zawacki, L. X. & Jannasch, H. W. *Deep Sea Res.* **37**, 695–710 (1990).

The Pisco Formation encompasses Mio-Pliocene marine transgressive sediments with abundant marine vertebrate fossils that accumulated in the shallow, nearshore waters of the Sacaco and Ica Basins of southern Peru^{1,2}. Major assemblages are assigned to seven vertebrate horizons of early-Middle Miocene to late-Early Pliocene age^{6,8}. The highly specialized odontocete described here was discovered in the Sud-Sacaco (SAS) horizon⁶ of earliest Pliocene age (about 5 Myr). It was found in the Sud-Sacaco locality⁹, situated about 0.5 km west of the panamerican highway, some 541 km south of Lima.

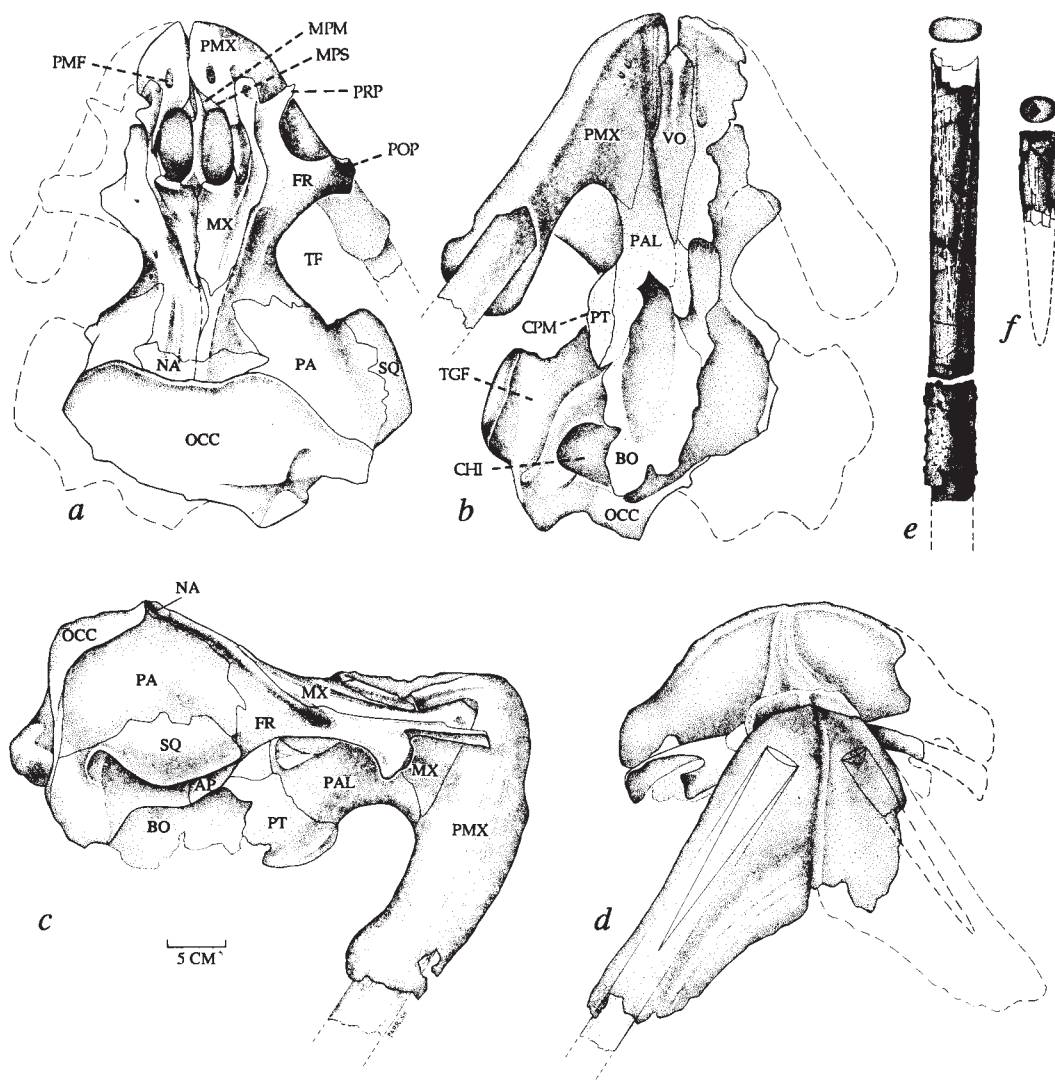
Superfamily Delphinoidea
Family Odobenocetopsidae new
Odobenocetops peruvianus gen. et sp. nov.

Holotype. The only specimen is an incomplete skull lacking much of the left side and all of the auditory region (Fig. 1) (U.S. National Museum of Natural History, Washington DC, USA USNM 460306).

Etymology. *Odon*, tooth (Greek); *baino*, walk (Greek); *cetus*, whale (Latin) and *ops*, like (Greek): 'the cetacean which seems to walk on its teeth', in reference to its feeding position (see below) and also to its similarity to the walrus (*Odobenus*); *peruvianus*, from Peru.

Diagnosis. Delphinoid cetacean (see Fig. 2) characterized by: loss of elongated cetacean rostrum and concomitant probable absence of melon; development of large ventrally directed premaxillary alveolar processes housing straight tusks; right tusk longer than 55 cm, cylindrical, with oval section and a long, open pulp cavity 23 cm deep; left tusk probably unerupted and probably no longer than 20 cm with a short, conical pulp cavity 1.5 cm deep; important muscular insertion and numerous neurovascular foramina on anterior side of premaxillae suggesting presence of strong upper lip and possible vibrissae (Fig. 3); regression in telescoping of the skull by forward migration of nares and anterior withdrawal of maxillae and frontals; nasals on the vertex of the skull, contacting occipital and right maxilla,

FIG. 1 The type skull of *Odobenocetops peruvianus* (USNM 460306) in dorsal (a), ventral (b), lateral (c) and anterior (d) aspects; (e) right premaxillary tusk in lateral view (bottom) and proximal view (top); (f), left premaxillary tusk in lateral view (bottom) and proximal view (top). Unpreserved parts are indicated by broken lines. In a, the dorsal view shows the reversal of the telescoping of the skull which drags anteriorly the narial fossae, the orbits, the frontals and the maxillae. As a consequence of this anterior withdrawal, the frontals do not cover the parietal and the temporal fossa as is observed in the other Delphinoidea. In the posterior part of the skull frontal and maxillae are reduced to narrow stripes connected to the nasals. The posterior extremity of the right maxilla is missing but its articular surface with the frontal is clearly visible on the specimen and the dotted lines indicate that it was contacting the right nasal. The nasals have remained posteriorly on the vertex of the skull, in a classical position for a delphinoid. Anteriorly, the loss of the rostrum allows no space for a melon and consequently that organ was probably lost in *Odobenocetops*. The position of the orbits suggests dorsal binocular vision. In c the lateral view of the premaxilla shows an elevated anteroventral edge (also visible in d) indicating the presence of strong ligament insertion and/or rough tissue which could have served as protection for the lip when foraging on the bottom. In d the tusks and their conical pulp cavities are represented as if they and the ventral bone were transparent; the small size of the left tusk as well as its shallow pulp cavity indicates that, as in most adult female dugongs and the right tusk of the male narwhal²⁰ the



tooth was unerupted but could probably have a slow growth in the alveolus. AP, alisphenoid; BO, basioccipital; CHI, cranial hiatus; CPM, rounded crest for the pterygoid muscle; FR, frontal; NA, nasal; MPM, medial portion of the maxilla; MPS, medial premaxilla-maxilla suture; MX, maxilla; OCC, occipital; PAL, palatine; PA, parietal; PMF, premaxillary foramen; PMX, premaxilla; POP, postorbital process; PRP, preorbital process; PT, pterygoid; SQ, squamosal; TF, temporal fossa; TGF, trough-like glenoid fossa; VO, vomer.

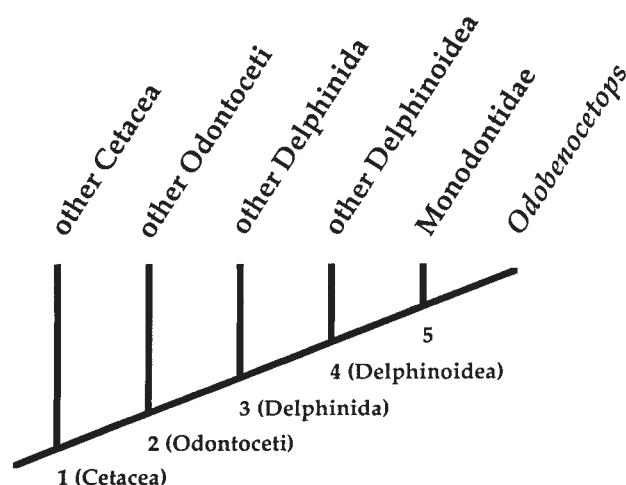


FIG. 2 Affinities of *Odobenocetops peruvianus*. The cladogram is based on the features preserved in the holotype (the only known specimen) of *O. peruvianus*. Consequently, several synapomorphies concerning the auditory region or the postcranial skeleton (unknown in *Odobenocetops*) are not considered here. The first four branchings designated by 'other + taxon name' do not necessarily represent monophyletic groups. Node 1; Cetacea: large overhanging supraorbital processes of frontals; narial openings facing dorsally²¹; peribullary and pterygoid air sinuses (expansions of the middle ear sinus) and lateral laminae of pterygoids (this feature is absent in *Pakicetus* (the oldest known cetacean), but is known exclusively in all other Cetacea, and therefore demonstrates the cetacean affinities of *Odobenocetops*)^{9,22}. Node 2, Odontoceti: maxillae partially covering supraorbital process of the frontals and expanded backwards far behind them^{23,24}; large premaxillary foramina²²; asymmetry of premaxillae, maxillae and frontals (asymmetry is not present in all odontocetes but it is known exclusively in this group and therefore demonstrates the odontocete affinities of *Odobenocetops*)^{23,24}; nasal passages subvertical; dorsoventrally expanded pterygoid sinuses. Node 3, Delphinida²⁵: palatine invaded posteriorly by the pterygoid air sinus and divided into medial and lateral laminae^{9,25}; formation of a cranial hiatus dorsal to petrosal by enlargement and fusion of the posterior and medium lacerate foramina^{22,24}. Node 4, Delphinoidea: medial portions of the maxillae inserted between the premaxillae at the posterior extremity of the rostral gutter and at the anteromedial border of the narial fossa (medial pmx-mx suture)^{9,25}. (This feature is not present in all Delphinoidea but it is known in this group only and therefore demonstrates the delphinoid affinities of *Odobenocetops*.) Node 5, Monodontidae + *Odobenocetops*: posterior extension of the lateral lamina of the palatine floors the optic gutter and contacts the frontal²⁵; alisphenoid and adjacent portion of the squamosal thickened medial to the zygomatic process and lateral to the foramen ovale²⁵; temporal fossa anteroposteriorly elongated; orbit migrated anteriorly.

laying on frontals but separated from mesethmoid which followed the bony nares in their forward migration; large temporal fossa opened dorsally; large dorsal exposure of parietals and concomitant development of temporalis muscle origin; orbits large, facing dorsally and not laterally as in other odontocetes; large pre- and postorbital processes expanded anterolaterally; maxillae articulating behind the nares; large, deep and vaulted palate; maxillae excluded from palate but part of the lateral wall of the skull; no maxillary teeth; no jugal; vomer very large and lanceolate; lateral lamina of the palatine flooring the anteroposteriorly oriented optic gutter; pterygoid with a flattened subhorizontal hamular process and large rounded lateral crest for origin of pterygoid muscle; ventral surface of zygomatic process of the squamosal forming a large and deep anteroposteriorly elongated trough-like glenoid fossa; large cranial hiatus; alisphenoid and squamosal strongly thickened on the anterolateral region of the cranial hiatus.

Odobenocetops is a delphinoid odontocete cetacean (Fig. 2). It constitutes the sister group of the Monodontidae (living narwhal and beluga) and its exceptional autapomorphies (loss of

the rostrum, alteration of the telescoping of the skull, ventrally directed tusks, anterior migration of the nares, and dorsal opening of the temporal fossa) warrant its position in a new family, the *Odobenocetopsidae*.

Odobenocetops and the living walrus show several morphological convergences: a large and deep palate; strong development of a wide, blunt snout which is highly vascularized and innervated with strong muscular insertions, suggesting the presence of a powerful, tactile upper lip; presence of tusks, and reduction of the maxillary dentition (post-canine maxillary teeth are absent in some fossil walruses¹⁰). The living walrus feeds mainly on benthic invertebrates, sucking out the siphon and foot of bivalve shells using the tongue as a 'piston' powered by very large lingual retractors and depressors. Their mouth works like a powerful vacuum pump^{11,12}. In *Odobenocetops* a similar feeding adaptation can be hypothesized. Furthermore, the orientation and shape of the glenoid fossa permitted anteroposterior movement of the mandible, which may have acted jointly with the tongue in suction. It is likely that a strong temporalis muscle worked with the masseter, tongue and throat muscles in mandibular retraction whereas the pterygoid muscles would have participated in protraction. The presence of a large (for a cetacean) temporal muscle origin expanded to the dorsal surface of the braincase, on the parietal, would be the result of anterior and medial withdrawal of the frontals and maxillae (this muscle is restricted to the temporal fossa in other odontocetes).

The melon is a characteristic odontocete organ involved in echolocation¹³. In *Odobenocetops*, the loss of this organ is related to the reduction of the rostrum and consequently echolocation was probably absent or very reduced.

It is likely that *Odobenocetops* fed upon benthic invertebrate fauna (bivalve and gastropod mollusks^{6,14} and/or crustaceans¹⁵), abundant in the Pisco Formation at Sacaco. When searching for food on the bottom with the body in an oblique position, the dorsally oriented orbit would have eventually provided a fairly good anterodorsal binocular vision and compensated for the poor echolocation ability. As in the living walrus, the tusks possibly served a primarily social function, and were not used as digging implements^{11,16}.

Although suction feeding has been increasingly recognized among odontocetes^{17,19}, the extreme modifications of *O. peruvianus* are unique among the Cetacea and suggest concomitant expansion of the adaptive breadth of the order. *Odobenocetops* occupied, in the Southern Hemisphere, an ecological niche otherwise unknown among odontocetes that was similar to that of the odobenine walruses which were widespread in the northern hemisphere during the Pliocene. □

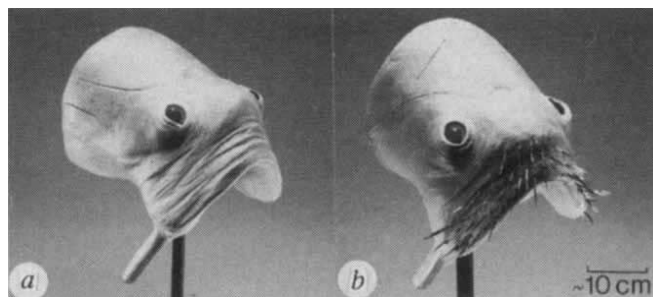


FIG. 3 Two possible reconstructions of the head of *Odobenocetops peruvianus*. a, Hypothesis without vibrissae, a 'cetacean-looking' interpretation; b, hypothesis with vibrissae, a 'walrus-looking' and more speculative interpretation. (Vibrissae are sometimes present in juvenile odontocetes²⁶, but otherwise vestigial in cetaceans except in the snout of the adult amazon dolphin²⁷.) Vibrissae, if actually present, would have had a tactile role when foraging for molluscs on the bottom as in the living walrus^{28,29}. Sculptures by M. Parrish, Department of Paleobiology, National Museum of Natural History.

Received 7 June; accepted 26 July 1993.

- Hoffstetter, R. G. *r. hebd. Acad. Sci. Paris sér. D* **287**, 1273–1276 (1968).
- Marocco, R. & de Muizon, C. *Géodynamique* **3**, 3–19 (1988).
- Marocco, R. & de Muizon, C. *Bull. Inst. Fr. Ét. Andines* **27**, 105–117 (1988).
- Dunbar, R. B., Marty, R. C. & Baker, P. A. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **77**, 235–261 (1990).
- Pilleri, G. *Beiträge zur Paläontologie der Cetaceen Perus* (Hirnanatomisches Institut, Ostermündigen, Bern, Switzerland) **I**, 1–233 (1989); **II**, 1–248 (1990).
- de Muizon, C. & DeVries, T. J. *Geol. Rdsch.* **74**, 547–563 (1985).
- de Muizon, C. *Éd. Rech. gr. Civilisations, mém.* **6**, 1–150 (*Trav. Inst. Fr. Ét. Andines* **22**) (1981).
- de Muizon, C. *Éd. Rech. Civilisations, mém.* **78**, 1–244 (*Trav. Inst. Fr. Ét. Andines* **42**) (1988).
- de Muizon, C. *Éd. Rech. Civilisations, mém.* **50**, 1–188 (*Trav. Inst. Fr. Ét. Andines* **27**) (1984).
- Deméré, T. A. J. *Paleont.* **10** (suppl.), 20A (abstract 48), (1990).
- Fay, F. H. N. *Am. Fauna* **74**, 1–279 (1982).
- Kastelein, R. A. & Mosterd, P. *Aquatic Mammals* **15**, 3–5 (1989).
- Mead, J. G. *Smithson. Contr. Zool.* **207**, 1–72 (1975).
- DeVries, T. J. in *Cenozoic Geology of the Pisco Basin Regional IGCP156 Field Workshop Genesis of Cenozoic Phosphorites and Associated Organic-rich Sediments: Peruvian Continental Margin* (eds Dunbar, R. B. & Baker, P. A.) 127–134 (Lima, Peru, 1988).
- Carriol, R. P., de Muizon, C. & Secrétan, S. *Annls Paleont.* **73**, 137–164 (1987).
- Oliver, J. S., Slattery, P. N., O'Connor, E. F. & Lowry, L. F. *Fish. Bull. F.A.O.* **81**, 501–512 (1983).
- Ray, C. in *Whales Dolphins and Porpoises* (ed. Norris, K. S.) Part IV, 671 (University of California Press, Los Angeles, 1966).
- Heyning, J. E. & Mead, J. G. *9th Biennial Conf. Biol. Marine Mammals*, Chicago 5–9 Dec. 1991, Abstracts p. 33.
- Werth, A. J. *9th Biennial Conf. Biol. Marine Mammals*, Chicago 5–9 Dec. 1991, Abstracts p. 73.
- Perrin, W. F. & Myrick, A. C. *Jr Rep. Int. Whal. Comm. Spec. Iss.* **3**, 1–229 (1980).
- Bourdelle, E. & Grassé, P.-P. in *Traité de Zoologie* (Masson Eds, Paris, 1955), **27**, 341–350.
- Barnes, L. G. in *The Bottlenose Dolphin* (eds Leatherwood S. & Reeves, R.) 3–26 (Academic, New York, 1990).
- Heyning, J. E. *Contr. Sci Nat. Hist. Mus Los Angeles County* **405**, 1–64 (1989).
- de Muizon, C. *Proc. Ser. San Diego Soc. Nat. Hist.* (in the press).
- de Muizon, C. *Annls Paleont.* **74**, 115–183 (1988).
- Brownell, R. L. in *Handbook of Marine Mammals Vol. 4* (eds Ridgway, S. H. & Harrison, R.) 45–67 (Academic, London, 1989).
- Robin, R. C. & da Silva, V. M. F. in *Handbook of Marine Mammals Vol. 4* (eds Ridgway, S. H. & Harrison, R.) 1–23 (Academic, London, 1989).
- Kastelein, R. A. & van Gaalen, M. A. *Aquatic Mammals* **14**, 123–133 (1988).
- Kastelein, R. A., Stevens, S. & Mosterd, P. *Aquatic Mammals* **16**, 78–87 (1990).

ACKNOWLEDGEMENTS. I thank C. E. Ray who invited me to study the specimen under his care, helped and encouraged me, and read the manuscript, C. Martin who found the specimen; P. L. Larson, who collected it and donated it to the USNM. P. Kroehler prepared the specimen. Illustrations are by M. Parrish; photographs are by C. Clark. L. G. Barnes, R. Bernor, A. Berta, D. J. Bohaska, T. A. Deméré, T. J. DeVries, D. P. Domning, R. E. Fordyce, J. E. Heyning, P. L. Larson and J. G. Mead gave valuable insights. The research was supported by a Short term Visitor Award of the Smithsonian Institution at Washington DC in 1992 and by funds from the Centre National de la Recherche Scientifique, Paris, France in 1991.

Cope's rule, the island rule and the scaling of mammalian population density

John Damuth

Department of Biological Sciences, University of California, Santa Barbara, California 93106, USA

COPE'S rule—the generalization that animal taxa tend to evolve toward larger body size—suggests that there are widespread net selective advantages to being large^{1–4}. Size–abundance relationships within bird^{5–7} and desert rodent⁴ guilds show that larger species usually do control more energy locally, and thus maintain larger populations than expected for their body size, implying that larger individuals are relatively better at obtaining and using local resources. But we report here results that show that this is not generally the case among mammal species. Within dietary groups containing only small species, larger species usually do better, but within those that contain the largest mammals, small species tend to control more energy. This suggests that in mammals there is an optimum body size for energy acquisition at about 1 kg. Thus, net adaptive advantages of large individuals for resource control cannot be used as a general explanation for evolutionary size increase in mammals, although other proposed explanations for Cope's rule are unaffected^{8–10}. Instead, these results suggest a partial explanation for another widespread ecotypic pattern, the 'island rule': that on islands, small mammal species evolve to larger size and large species to smaller size^{11–13}. If on an island a species' usual competitors and predators are absent, it should often tend to evolve toward the optimum body size, and the adaptive advantages of doing so would be greatest for populations starting at body-size extremes.

There are several commonly used measures of animal abundance^{14–16}, and each, when related to body size, is appropriate for addressing a different set of questions about ecological and macroevolutionary relationships¹⁶. 'Ecological' population densities are those measured over an area of habitat actually used by a species. Comparing them reveals little directly about the relative distribution of species of different size within an ecologically heterogeneous region or ecosystem. However, the scaling of ecological densities does allow assessment of the relative degrees to which species of different sizes have adapted to extract energy from their local environment. This is the question usually asked in the macroevolutionary context of Cope's rule.

For a species to maintain a large population relative to its body size, and thus to dominate energy flow in its local habitat, it must be composed of individuals that are better than those of other species at obtaining or using energy. For example, larger individuals may be better at interspecific competition and may expend less energy in thermoregulation and predator avoidance⁴. If local populations of large species are usually energetically dominant to small species, this implies that larger size generally confers net advantages in energy control, and provides a possible selective explanation for Cope's rule⁴. Because metabolic requirements of individuals scale with body mass (M) roughly as $M^{0.75}$ (refs 17, 18), populations of different-sized species are using (on average) the same amount of trophic energy when their ecological densities scale reciprocally, as $M^{-0.75}$ (ref. 19). If the exponent for density scaling is greater than -0.75 , the larger species in the sample are, on average, controlling more energy. If the exponent is less than -0.75 , the smaller species are doing better on average. When the data are log-transformed, the exponents equal the slopes of linear regressions of density on body mass (see Fig. 1). If Cope's rule is explained by large size conferring individual adaptive advantages, the expectation is that ecological densities will tend to scale across species with regression slopes greater than -0.75 .

In samples pooling data from a wide range of body sizes, habitats and dietary groups, ecological densities of mammals consistently scale with slopes near -0.75 (refs 19–21). This means that, overall, large mammal species do not have the ability to extract energy from their local habitats at a different rate from small mammal species. But analyses within vertebrate feeding guilds and dietary groups suggest that regression slopes at this level of analysis are usually less than -0.75 , and sometimes are positive^{4, 7, 22}. This implies that, relative to their presumed major competitors, individuals of larger species usually do control more trophic energy. However, only a small and ecologically restricted set of mammalian dietary groups have been analysed so far^{4, 22}.

Ecological densities, mean adult body mass and habitat and dietary data for 557 species of terrestrial, non-volant mammals were obtained from the literature (full data set available from the author on request). This represents an update and expansion of a previous compilation²⁰ and the addition of 90 species for which data have become available. Each species was assigned to one of nine diet categories (corresponding to feeding guilds), one of three habitat types, and a geographic/biotic region (Table 1). Regressions of density on body mass were calculated on the logarithmically transformed variables for each dietary group in each habitat-type and region (Fig. 1 and Table 1). Not all diet

TABLE 1 Regression statistics for mammalian dietary groups

Diet	Habitat	Region	Slope	Int.	N	R ²	P	RMSE	σ^2 Mass	Mean mass	Min. mass	Max. mass
CA	C	E/NA	-1.30	4.46	8	0.76	0.0050	0.5354	0.4503	3.944	2.94	4.86
CA	C	NEOT	-1.29	4.95	5	0.78	0.0473	0.3530	0.1991	4.466	3.87	4.93
CA	C	TRAS	-1.36	5.61	3	0.99	0.0769	0.1665	0.5142	4.339	3.70	5.11
CA	O	E/NA	-0.66	2.57	4	0.83	0.0889	0.4547	1.5350	2.971	1.90	4.15
CA	O	AFRC	-0.64	2.23	12	0.36	0.0393	0.6171	0.4804	3.989	2.72	5.18
FG	C	E/NA	-0.28	3.28	17	0.11	0.1947	0.4275	0.2531	1.812	1.30	2.83
FG	C	NEOT	-0.85	4.18	25	0.53	0.0001	0.5369	0.4273	2.728	1.59	3.91†
FG	C	TRAS	-1.00	4.43	4	0.80	0.1031	0.3315	0.3021	2.688	2.00	3.23
FG	O	E/NA	0.085	2.88	26	0.0063	0.7004	0.4871	0.1969	1.688	0.85	2.55
FG	O	AFRC	-0.37	3.40	12	0.20	0.1472	0.3621	0.2197	1.583	0.70	2.41
FG	D	E/NA	-0.054	2.81	13	0.002	0.8715	0.5820	0.2610	1.604	0.87	2.42*
FH	C	AFRC	-0.70	4.00	9	0.66	0.0077	0.5009	0.8732	3.486	1.85	4.65
FH	C	MADR	-1.33	6.81	6	0.58	0.0802	0.4670	0.1343	3.444	3.03	4.10
FH	C	TRAS	-0.95	5.00	16	0.24	0.0540	0.4221	0.05876	3.911	3.71	4.72
FH	O	SATM	0.42	1.92	6	0.06	0.6376	0.7790	0.1771	1.977	1.59	2.60
FO	C	E/NA	-1.04	4.93	8	0.90	0.0003	0.3797	1.0783	2.087	1.26	4.03
FO	C	AFRC	-0.47	3.06	16	0.38	0.0113	0.4815	0.5888	3.014	1.58	3.90
FO	C	NEOT	-0.60	3.13	38	0.38	0.0001	0.4552	0.3416	2.788	1.54	3.85*
FO	C	TRAS	-0.61	3.37	10	0.13	0.3012	0.9381	0.3213	2.269	1.73	3.70
FO	O	AFRC	-0.78	3.80	15	0.82	0.0001	0.5796	2.3672	2.853	0.81	4.82
FO	O	NEOT	-0.81	4.02	6	0.63	0.0602	0.6475	0.8713	1.954	1.38	3.85
HZ	C	E/NA	-0.88	5.26	3	0.99	0.0550	0.2324	4.6836	3.558	1.69	5.93
HZ	O	E/NA	-0.94	5.15	11	0.82	0.0001	0.5874	1.5745	2.421	1.43	5.74
HZ	O	AFRC	-0.87	4.90	24	0.82	0.0001	0.6227	2.2537	4.227	1.85	6.35
HZ	O	NEOT	-0.13	2.23	3	0.31	0.6223	0.3721	1.9847	3.089	1.70	4.52
MX	C	TRAS	-0.45	2.47	6	0.83	0.0122	0.1632	0.4991	5.316	4.15	6.26
MX	O	AFRC	-0.30	1.68	17	0.10	0.2075	0.5749	0.3982	4.749	3.91	6.46
MX	O	TRAS	-0.37	2.69	5	0.23	0.4126	0.5492	0.5070	5.037	4.47	6.10
HB	C	E/NA	-0.85	4.52	5	0.65	0.0980	0.7120	0.9783	4.535	3.13	5.61
HB	C	AFRC	-1.00	5.24	10	0.63	0.0063	0.5109	0.3930	4.160	3.39	5.40
HB	C	NEOT	-1.21	6.48	7	0.96	0.0001	0.2416	0.7506	4.148	3.38	5.48*
HB	C	TRAS	-1.09	5.74	8	0.74	0.0061	0.7352	1.1188	4.346	3.43	6.05
HB	O	E/NA	-0.94	5.29	7	0.70	0.0184	0.8039	1.4524	3.181	1.54	4.95
HB	O	AFRC	-0.52	2.83	8	0.44	0.0731	0.6053	0.9244	4.689	3.48	5.98
IO	C	E/NA	-0.63	3.13	8	0.70	0.0101	0.5210	1.3219	1.238	0.62	4.04
IO	C	NEOT	-0.55	2.92	24	0.40	0.0010	0.4719	0.4670	2.345	1.25	3.64*
IO	O	E/NA	-0.87	4.08	7	0.65	0.0288	0.8181	1.3630	2.083	0.85	3.96
IO	O	AFRC	-0.82	3.57	13	0.74	0.0002	0.5165	1.0014	1.808	0.58	3.57
MY	C	NEOT	-0.81	3.36	6	0.75	0.0266	0.3624	0.4730	3.744	2.60	4.60*

Dietary classification follows Eisenberg^{22,28}, with addition of a mixed browser/grazer category for herbivores²⁹: CA, carnivore; MY, myrmecophage; IO, insectivore-omnivore; FO, frugivore-omnivore; FG, frugivore-granivore; FH, frugivore-herbivore; HB, herbivore-browser; MX, mixed browser/grazer; HZ, herbivore-grazer. Habitat: C, closed (forest); O, open (savanna, grassland; tundra excluded); D, desert. Region: E/NA, non-tropical North America and Eurasia; NEOT, neotropics; SATM, temperate South America; TRAS, tropical Asia; AFRC, tropical Africa; MADR, Madagascar. Slope, int: slope and intercept of least-squares regression of $\log_{10}$ population density (km^{-2}) on $\log_{10}$ body mass (g). N, sample size. R², coefficient of determination. P, probability of null hypothesis that slope = 0. RMSE, root mean square error (standard error of the estimate); σ^2 Mass, variance of $\log_{10}$ body mass. The mean of all 39 slopes is -0.71 (s.e. = 0.064).

* Slope agrees qualitatively (with respect to whether large or small species are using more energy) with previously published studies^{4,22}.

† Slope does not agree qualitatively with previous study²².

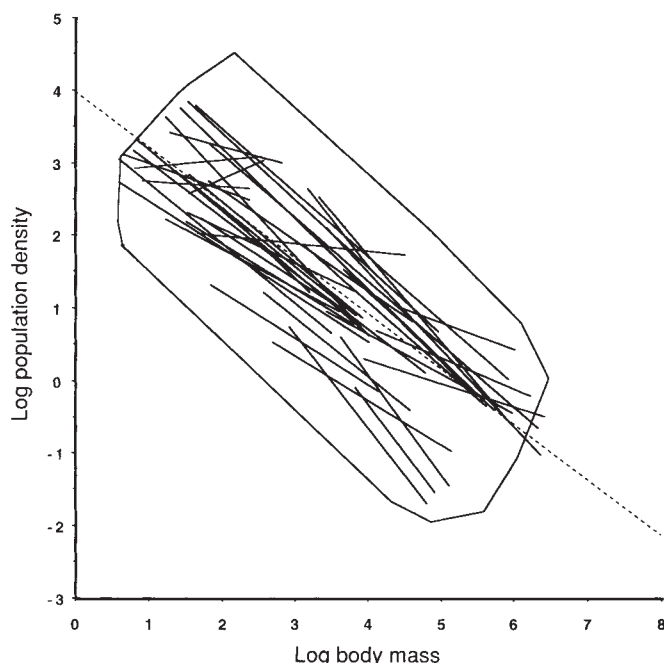
categories in each region were represented by sufficient numbers of species (>2) for a regression. Potential regressions were also rejected if the body-mass range spanned less than one order of magnitude.

The regression slopes for 21 of the 39 resulting dietary groups are less than -0.75, and only two slopes are positive. The null hypothesis that overall the slopes are symmetrical about -0.75 cannot be rejected (binomial test, $P=0.75$). Overall, slope does not depend upon diet (Kruskal-Wallis, $P=0.14$). Slopes are, however, negatively correlated with the size of the species within a dietary group, expressed either as the midpoint of the group's range in body mass ($P=0.02$, Spearman rank correlation $S\rho=-0.38$) or as the maximum body mass of the group ($P=0.03$, $S\rho=-0.34$). Slopes are not correlated with the range of body mass ($P=0.81$, $S\rho=-0.04$) or the variance of body mass ($P=0.30$, $S\rho=-0.17$), indicating that error-variance artefacts²³ are not causing this negative correlation. In five out of six cases where slopes from this analysis are directly comparable to those of previous analyses the results are qualitatively the same (Table 1).

Slopes within mammalian dietary groups are thus usually negative and consistent with earlier analyses that implied that large species of mammals do not have an advantage in energy capture. There is no empirical basis in abundance scaling to suggest that selection within species should generally favour individuals of large size, or that populations of larger mammal species should usually be ecologically dominant to those of smaller species. In this respect mammal guilds may be different from bird guilds^{4,7}, but it is worth noting that birds are one of the few taxa within which Cope's rule has not been observed⁸. Large body size may nevertheless still be associated with other kinds of macroevolutionary advantages^{9,24}. In view of the results reported here, explanations for Cope's rule that do not depend upon individual adaptive advantages of large size^{8,10} are the more plausible.

The negative correlation between slope and the size of species in a dietary group reflects an important qualitative difference between guilds of small species and those that contain large species. Dietary groups whose largest members are ≤ 10 kg tend to have slopes greater than -0.75 (larger species control

FIG. 1 Ecological population density (km^{-2}) regressed on body mass (g) for mammal species. The polygon encloses the entire sample of 557 mammal species, and the dotted line shows the ordinary least-squares (OLS) regression of this pooled sample ($\log \text{ density} = -0.77 \log \text{ mass} + 3.98$; $R^2 = 0.635$; $\text{RMSE} = 0.763$; $\sigma^2 = 1.729$; $P = 0.001$). In 7 cases separate density estimates were available for a single species in two habitats or regions, so $n = 564$ for the pooled regression. The lines within the polygon show the 39 OLS regressions for mammalian dietary groups (Table 1; 424 species, total), drawn in each case only over the range of the data. (There can be considerable scatter about the regressions. In the 3 cases where the drawn regression lines cross the polygon boundary, the terminal data points (not shown) lie off the line but still within the polygon.) Elevations of the lines are strongly related to diet (Kruskal-Wallis, $P = 0.0007$, for density evaluated at the mid-point of the body-mass range), with carnivores and myrmecophages showing the lowest average density values for their size. The OLS regression line will represent faithfully the true functional relationship of the variables only if population density estimates are subject to many more times the uncontrolled statistical 'error' than are body mass estimates³⁰⁻³². The current data satisfy this condition. These data incorporate substantial random deviations from the true linear relationship between size and density in each case that are primarily the result of factors that affect population density values independently of body size. These factors include differences in quality of habitat, temporal variation in habitat quality or social behaviour, short-term population demographic history, and presence of competitors or predators. Furthermore, the variance associated with measurement error for population density estimates is much larger than that for the mean body mass of a species^{33,34}. Under these conditions, OLS regression is appropriate and gives less biased results than do other line-fitting techniques such as major axis or reduced major axis^{31,32}, which would give more steeply negative slopes in most of the above cases. The dietary groups are composed of different combinations of species from a variety of mammalian higher taxa, with different phylogenetic histories. Phylogenetic relationships do not affect the analysis, because population density is an instantaneous ecological characteristic that does not coevolve with body mass³⁵. Nevertheless, data points for individual species are probably not entirely statistically independent³⁶, resulting in an overestimate of the number of degrees of freedom involved in each regression. But none of the nonparametric statistical tests used here use information on the sample sizes or degrees of freedom of the regressions. This analysis does not directly compare the energy control of populations



belonging to the same feeding guilds in local communities. Too few mammal communities have been fully censused. Rather, the research design corresponds to a scheme where each species is sampled independently from one or more of the communities that it inhabits in a region, and the independently derived mean densities are compared within dietary categories corresponding to feeding guilds. Computer simulations (J.D., unpublished) show that over a wide range of conditions representative of the variation seen in natural faunas the slope values derived from such pooled data faithfully represent the true distribution of guild slopes within communities. The question relevant to Cope's rule is whether the average value for guilds shows a slope greater than -0.75 . The slopes for dietary groups analysed here can answer this question statistically.

more energy), whereas those that contain species >10 kg usually have slopes less than -0.75 (smaller species control more energy). In these data, the only major exceptions appear to be open-habitat carnivores and mixed-feeding browser/grazers (perhaps an unusually heterogeneous category). Figure 1 suggests that it is at medium size (~ 1 kg) that species control most energy in most dietary groups. This is consistent with suggested physiological and reproductive advantages of intermediate body size in mammals²⁵.

This pattern provides a possible explanation for the so-called 'island rule'¹²: relative to mainland populations, there is a widespread tendency for small mammals to evolve to larger size on islands and large mammals to evolve to smaller size; there is no particular trend in size change for medium-sized mammals¹³. Island populations frequently experience a decrease in the number of both competitor and predator species^{13,26,27}. Under these conditions we would expect an island population to be free to evolve toward the body size most advantageous for exploiting resources in its diet category. For species not near the intermediate, optimum size for energy control, this will mean directional evolutionary change in body size conforming to the island rule. Other factors peculiar to island habitats may also contribute to the observed size changes^{12-14,26,27}, but part of the explanation for the island rule is likely to be the general advantage of medium size that is observed even in continental faunas. □

- Newell, N. D. *Evolution* **3**, 103-124 (1949).
- Gould, S. J. *Biol. Rev.* **41**, 587-640 (1966).
- Brown, J. H. & Maurer, B. A. *Nature* **324**, 248-250 (1986).
- Nee, S., Read, A. F., Greenwood, J. J. D. & Harvey, P. H. *Nature* **351**, 312-313 (1991).
- Juanes, F. *Am. Nat.* **128**, 921-929 (1986).
- Cotgreave, P. & Harvey, P. H. *Functional Ecol.* **6**, 248-256 (1992).
- Stanley, S. M. *Evolution* **27**, 1-26 (1973).
- Van Valen, L. M. *Evolution* **29**, 87-94 (1975).
- Gould, S. J. *J. Paleont.* **62**, 319-329 (1988).
- Foster, J. B. *Nature* **202**, 234-235 (1964).
- Van Valen, L. M. *Evol. Theory* **1**, 31-49 (1973).
- Lomolino, M. V. *Am. Nat.* **125**, 310-316 (1985).
- Lawton, J. H. *Oikos* **55**, 429-434 (1989).
- Lewontin, R. C. & Levins, R. *Am. Nat.* **134**, 513-524 (1989).
- Damuth, J. *Nature* **351**, 268-269 (1991).
- Kleiber, M. *The Fire of Life* 2nd edn (Krieger, New York, 1975).
- Nagy, K. A. *Ecol. Monogr.* **57**, 111-128 (1987).
- Damuth, J. *Nature* **209**, 699-700 (1981).
- Damuth, J. *Biol. J. Linn. Soc.* **31**, 193-246 (1987).
- Damuth, J. *Nature* **351**, 268-269 (1991).
- Robinson, J. G. & Redford, K. H. *Am. Nat.* **128**, 665-680 (1986).
- Pagel, M. D. & Harvey, P. H. *Am. Nat.* **132**, 344-359 (1988).
- du Toit, J. T. & Owen-Smith, N. *Am. Nat.* **133**, 736-740 (1989).
- Maiorana, V. C. in *Body Size in Mammalian Paleobiology* (eds Damuth, J. & MacFadden, B. J.) 69-102 (Cambridge Univ. Press, 1990).
- Sondaar, P. Y. in *Major Patterns of Vertebrate Evolution* (eds Hecht, M. K., Goody, P. C. & Hecht, B. M.) 671-707 (Plenum, New York, 1977).
- Heaney, L. R. *Evolution* **32**, 29-44 (1978).
- Eisenberg, J. F. *The Mammalian Radiations* (Univ. Chicago Press, 1981).
- Hofmann, R. R. & Stewart, D. R. *Mammalia* **36**, 226-240 (1972).
- Kuhry, B. & Marcus, L. F. *Syst. Zool.* **26**, 201-209 (1977).
- Riggs, D. S., Guarneri, J. A. & Adelman, S. *Life Sci.* **22**, 1305-1360 (1978).
- McArdle, B. H. *Can. J. Zool.* **66**, 2329-2339 (1988).
- Yabloukov, A. V. *Variability of Mammals* (Amerind, New Delhi, 1974).
- Caughley, G. *Analysis of Vertebrate Populations* (Wiley, London, 1977).
- Felsenstein, J. *Am. Nat.* **125**, 1-15 (1985).
- Ridley, M. J. *theor. Biol.* **136**, 361-364 (1989).

Received 20 April; accepted 28 July 1993.

1. Rensch, B. *Evolution above the Species Level* (Columbia Univ. Press, New York, 1959).

ACKNOWLEDGEMENTS. I thank S. J. Mazer for comments on the manuscript.

Visual pattern recognition in *Drosophila* involves retinotopic matching

Marcus Dill, Reinhard Wolf & Martin Helsenberg*

Theodor-Boveri-Institut für Biowissenschaften (Biozentrum),
Lehrstuhl für Genetik, Am Hubland, D-97074 Würzburg, Germany

* To whom correspondence should be addressed

HONEYBEES remember the shapes of flowers and are guided by visual landmarks on their foraging trips^{1,2}. How insects recognize visual patterns is poorly understood. Experiments suggest that they try to match retinotopically the incoming visual pattern with a previously stored memory image²⁻⁷. But bees can be conditioned to individual pattern parameters such as orientation of contours, colour or size^{2,8-11}. These and other results are difficult to reconcile with simple template matching. In such investigations, freely moving animals are observed; their behaviour and visual input, therefore, are not well known. Mostly, processing strategies are inferred from stimulus design. We have studied visual pattern recognition with tethered flies (*Drosophila melanogaster*) in a flight simulator and report here that flies store visual images at, or together with, fixed retinal positions and can retrieve them from there only⁵. Position invariance, an acknowledged property of human pattern recognition, may not exist as a primary mechanism in insects.

Flies were attached to a torque meter with their heads immobilized, and allowed to control with their yaw torque the angular velocity of a vertical drum surrounding them (Fig. 1). The internal surface of the drum was covered with either a $2 \times 180^\circ$ random-dot texture or four regularly spaced figures. This simulated free flight in the horizontal plane with infinitely distant landmarks (flight simulator)¹².

In the flight simulator, flies effectively avoid certain directions of flight relative to the panorama (Fig. 2b) if these are associated with a noxious stimulus such as heat from an infrared light-source. For experimental convenience, flies were heated whenever they fixated any location in two opposing quadrants of the

panorama, whereas heat was turned off as long as they fixated the intervening sectors. Opposing quadrants were marked by identical visual patterns (textures in Fig. 2). Flies still avoid patterns previously associated with reinforcement in subsequent tests without heat. This memory can last for more than 20 hours.

To ensure that the memory is indeed based on pattern recognition and not on spatial orientation alone, four identical patterns were presented in the arena. Flies still effectively avoid the heat during training but, as expected, fail in the test¹³. The experiment must be classified as operant conditioning. Flies need to be in command of the rotation of the drum during training. Mere pairing of the reinforcement with a moving visual pattern at certain angular positions does not affect subsequent pattern preferences¹³.

With the experimental flies fixed in space the retinotopic matching hypothesis can be tested much more rigorously than before. For retrieval of a learned pattern according to this mechanism, flies would be required to shift the pattern to the same retinal position in which they had stored it. Indeed, during the experiments with the random-dot texture, flies acquire individual preferences for particular orientations relative to the landmarks in the texture. This behaviour is apparent already in control flights without reinforcement (Fig. 2c, open circles) and is even more striking during the training periods in which the flies may receive punishment (Fig. 2c, filled circles).

In the flight simulator, flies are in command of horizontal pattern motion only. Therefore, flies would not be able to move the pattern to the same location if this had been vertically displaced between training and test. Such an experiment is shown in Fig. 3. During training the two pairs of patterns were presented at 15° (Fig. 3a; 4.5° and 1.5° in b and c, respectively) above or below the 'horizon' and are displaced before the test by 30° (9° , 3°) across the horizon to the corresponding height. (Patterns are handled in the same manner in the controls but are placed back into their previous positions).

A pattern shift of 3° , which is well below the mean divergence angle between adjacent visual elements ($\Delta\phi = 4.6^\circ$)^{14,15} and also below the mean half width of the angular sensitivity distribution of individual visual elements ($\Delta\rho = 5^\circ$)¹⁴, has little if any effect on the recognition of the memory template. Larger displace-

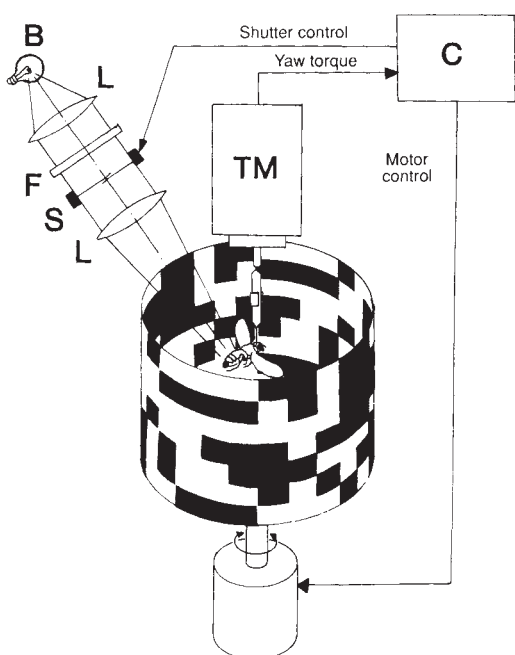


FIG. 1 Experimental design. The animal is tethered to a device transducing the fly's yaw torque into a DC voltage (TM, torque meter). This in turn is used to control (C, computer) the angular velocity of an arena surrounding the fly (coupling coefficient $-k = 11^\circ (\text{s } 10^{-10} \text{ N m})^{-1}$; for details of the apparatus see refs 13, 22). The cylindrical wall of the arena, which is illuminated from behind (mean luminance $\bar{l} = 10 \text{ cd m}^{-2}$), is covered by a random-dot texture (Figs 1 and 2) consisting of 12.9° wide, black and transparent squares in random sequence. The pattern between 0 and 180° is repeated in tandem from 180 to 360° yielding point symmetry at the position of the fly. Alternatively (Figs 3 and 4), the four quadrants of the cylinder are marked in their centre by isolated figures (T, triangle) as indicated, with opposing quadrants carrying identical patterns. Negative reinforcement is provided by heating the fly with a microscope lamp (Zeiss, 6V, 15W). The beam is passed through an infrared filter (F; Schott RG780, 3 mm), focused onto the fly from above by an $f = 60 \text{ mm}$ lens (L) and can be intercepted by a computer-controlled shutter (S). The shutter is open as long as the fly fixates any position in two opposing 'punished' quadrants; it is closed during fixation of the intervening quadrants. The position of the arena is continuously stored in the computer. A performance index (PI) is calculated for 2-min flight periods $(t_1 - t_2)/(t_1 + t_2)$, with t indicating the time the fly spent fixating the no-heat (t_1)- and heat (t_2)-associated quadrants. Between training and test, the drum position is 'randomized' (rotated at high speed and then stopped at an arbitrary position). In all experiments, half the flies are conditioned to avoid one pattern and the other half are conditioned using the alternative pattern to eliminate from the data any spontaneous pattern preferences and asymmetries in the apparatus. Statistical comparison of experimental groups: *U*-tests after Wilcoxon, Mann and Whitney. To test whether performance indices differ significantly from zero, individual data are transformed (arcsine) and subsequently evaluated using *t*-tests.

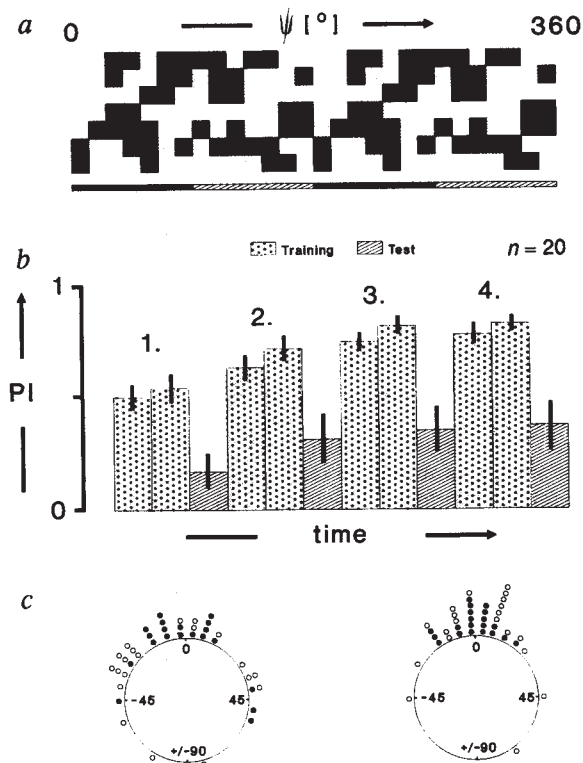


FIG. 2 Visual pattern learning in the flight simulator. Columns in *b* indicate the mean performance indices (PI) for consecutive 2-min periods from 20 flies (*n*). Half of the animals are conditioned to the underlined quadrants in *a*, the second half to the alternate ones. The arena is shifted to a random position before each test for acquired pattern preferences (dark shading) forcing the animals to use visual pattern discrimination for their choice. Five different random-dot patterns are used in the experiment to avoid special properties of single transparencies. Only one is shown in *a* with two identical quadrants underlined. In *b* the performance index of 0.37 ± 0.11 during the last test is significantly different from the index -0.08 ± 0.18 (not shown in the figure) reached by a control group of 10 animals tested for spontaneous preferences with the same random patterns without any heat reinforcement (*U*-test, $U=122$, $P<0.05$). Vertical bars are s.e.m.s. *c*, Mean flight directions for the first and third 4-min training periods are calculated for each fly. These directions are used as 'zero' references to which the individual mean flight directions (indicated as dots in polar plots) in the second and fourth training periods, respectively, are compared. All relative angular orientations $\psi > 180^\circ$ are transformed into ψ' -values by $\psi' = \psi - 180^\circ$. Mean orientations are calculated from ψ and ψ' -values. In the control flights without heat the same respective periods are used for the calculations.

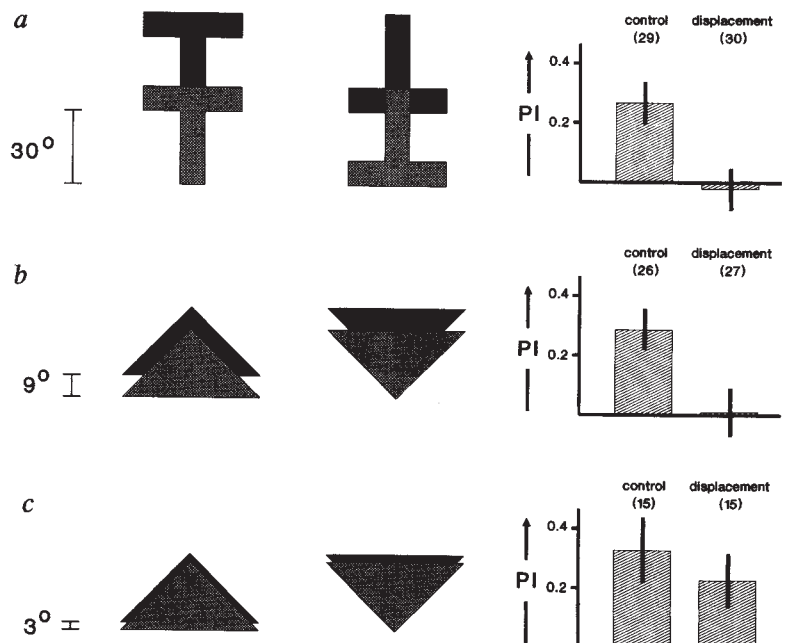
ments (9° , 30°), however, effectively prevent recognition, indicating that it depends on a very precise matching process.

If the retinal position of patterns is as crucial for their recognition as the above experiments suggest, flies might be able to discriminate patterns exclusively on the basis of their retinal position. As discussed above, with four identical patterns flies can avoid the heated sectors during training but not in the test. If, however, these patterns are presented at slightly different heights relative to each other they are distinguished in the test and the learning index is as high as with differently shaped patterns (Fig. 4). Apparently, 'retinal position' that continuously

changes in free flight is not only necessary but also sufficient for the fly to recognize a previously learned pattern.

Position invariance may not exist in insect pattern recognition. Our data strongly suggest that, for recognizing a memorized visual pattern, flies have to shift the actual image to the same location in the visual field where the image had been presented during the storage process. Also, in free flight, no interocular and vertical retinal transfer was found in bees^{2,3}. For pattern recognition in large visual systems, position invariance is not in dispute^{16,17}. But retinotopic matching may play a major role even in human visual perception, considering that no

FIG. 3 The effect of vertical displacement on pattern recognition. Flies are trained for 8 half-minute periods (each separated by one half-minute period without reinforcement) with patterns in one of two different heights as indicated on the left. The two heights are symmetrically positioned with respect to the fly's 'horizon'. Between training and test, the arena is removed in the dark and the pattern transparency either is exchanged to a new one containing identical patterns at an alternative height, or the same transparency is inserted (controls). In experiment *a*, upright and inverted Ts (vertical size 40°) are displaced by 30° , the triangles in *b* (vertical size 30°) by 9° . In *c* the triangles are shifted by only 3° . In each of the displacement and control groups (low $\rightarrow$ high; low $\rightarrow$ low; high $\rightarrow$ low; high $\rightarrow$ high) half of the flies are conditioned to avoid one pattern, the other half the alternative one. Differences between control and displacement groups are significant in *a*: *U*-test, $U=172$, $P<0.01$; and *b*: *U*=140, $P<0.05$. The difference in *c* is not significant ($P>0.3$). Half-minute measurements without reinforcement at the beginning of the experiment (not shown) did not detect any significant spontaneous pattern preferences (*t*-test: $P>0.3$ for all three experiments). Numbers above columns indicate number of flies tested. Vertical bars, s.e.m.s.



retinal transfer is observed in various kinds of perceptual learning^{18,19}.

This study does not directly address the question of whether templates are stored as 'eidetic' images⁴⁻⁷ or as sets of characteristic parameters^{2,8-11}. Although the results favour the former alternative they can also be reconciled with the latter if the animals use retinal position as a crucial parameter. By contrast, in many of the experiments that cannot be explained by simple retinotopic matching, perhaps insects extract individual pattern parameters by comparing different consecutively stored templates⁸⁻¹⁰.

The pattern position traces in the flight simulator reveal that the retinal image is in continuous motion except during the brief moments (<30 ms) of rotational inversion¹². Template formation may thus require pattern motion. Feature detection neurons specific for the orientation of moving bars or edges have recently been recorded in the dragonfly optic lobe²⁰. If the template is static as Srinivasan *et al.*²¹ have argued, it must be blurred unless exposure time is very short. The flight simulator may help to bridge the gap between free-flight behaviour and single unit recording. □

Received 5 July; accepted 3 September 1993.

1. Tinbergen, N. Z. *vergl. Physiol.* **15**, 305-334 (1932).
2. Wehner, R. in *Handbook of Sensory Physiology* Vol. VII/6C (ed. Autrum, H. J.) (Springer, Berlin, 1981).
3. Wehner, R. J. *comp. Physiol.* **77**, 256-277 (1972).
4. Cartwright, B. A. & Collett, T. S. *Nature* **295**, 560-564 (1982).
5. Cartwright, B. A. & Collett, T. S. J. *comp. Physiol.* **A161**, 521-543 (1983).
6. Gould, J. L. *Science* **227**, 1492-1494 (1985).
7. Collett, T. S. *Phil. Trans. R. Soc. B337*, 295-303 (1992).
8. Ronacher, B. *Biol. Cybern.* **32**, 63-75 (1979).
9. van Hateren, J. H., Srinivasan, M. V. & Wait, P. B. J. *comp. Physiol.* **A167**, 649-654 (1990).
10. Horridge, G. A., Zhang, S. W. & Lehrer, M. *Phil. Trans. R. Soc. B337*, 49-57 (1992).
11. Zhang, S. W., Srinivasan, M. V. & Horridge, G. A. *Proc. R. Soc. B249*, 55-61 (1992).
12. Heisenberg, M. & Wolf, R. in *Visual Motion and its Role in the Stabilization of Gaze* (eds Miles, F. A. & Wallmann, J.) 265-283 (Elsevier, Amsterdam, 1993).
13. Wolf, R. & Heisenberg, M. J. *comp. Physiol.* **A169**, 699-705 (1991).
14. Götz, K. G. *Kybernetik* **2**, 215-221 (1965).
15. Buchner, E. *Biol. Cybern.* **24**, 85-101 (1976).
16. Cronly-Dillon, J. R., Sutherland, N. S. & Wolfe, J. J. *exp. Neurol.* **15**, 455-462 (1966).
17. Myers, R. E. J. *comp. Physiol. Psychol.* **48**, 470-473 (1955).
18. Ramachandran, V. S. *Nature* **282**, 382-384 (1976).
19. Karni, A. & Sagi, D. *Proc. natn. Acad. Sci. U.S.A.* **88**, 4966-4970 (1991).

Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task

Shintaro Funahashi*, Matthew V. Chafee & Patricia S. Goldman-Rakic†

Section of Neurobiology, Yale University School of Medicine, New Haven, Connecticut 06510, USA

PATIENTS with damage to the dorsolateral prefrontal cortex are impaired on cognitive tasks such as the Wisconsin Card Sort Test¹, the Stroop Test² and an anti-saccade paradigm³, in which sensory-guided habitual responses must be suppressed in favour of conceptually or memory-guided responses. We report here recordings from prefrontal neurons in rhesus monkeys trained to perform a delayed anti-saccade task based on tests that have been used with humans³. Activity in the same prefrontal neurons was recorded across conditions when saccades were made toward a remembered target, and also when this prepotent response was suppressed and a saccade in the opposite direction required. Our findings show that most prefrontal neurons code the location of the visual stimulus in

*Present address: Faculty of Integrated Human Studies, Kyoto University, Sakyo-ku, Kyoto 606-01, Japan.

† To whom correspondence should be addressed.

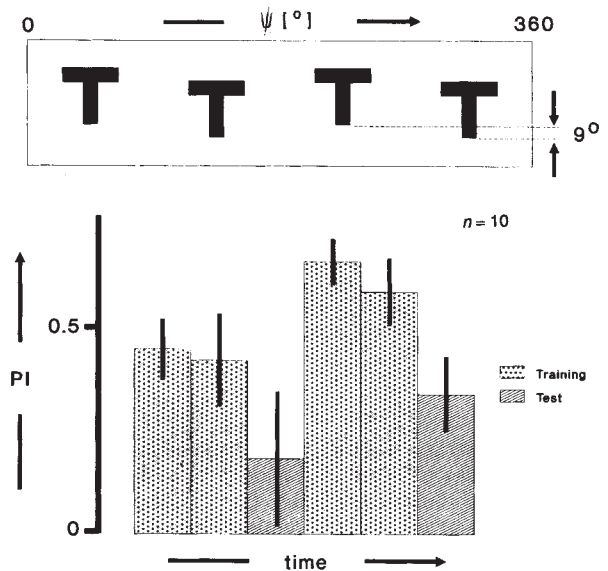


FIG. 4 Flies can discriminate identical figures (T) that are presented at slightly different heights (difference: 9°). Each column represents a 2-min training or test period (striped). Half of the flies are conditioned to avoid the upper figures, the other half the lower ones. Between training and test, the arena is shifted to a random angular position. The performance index (PI) of 0.37 ± 0.10 in the final test is significantly different from zero (*t*-test, $t=3.44$, $P<0.01$). Controls confirm that with four identical Ts at the same height no learned pattern preference can be detected after randomizing¹³. Vertical bars, s.e.m.s; *n*, total number of flies.

20. O'Carroll, D. *Nature* **362**, 541-543 (1993).

21. Srinivasan, M. V., Zhang, S. W. & Rolfe, B. *Nature* **362**, 539-540 (1993).

22. Heisenberg, M. & Wolf, R. J. *comp. Physiol.* **A163**, 373-388 (1988).

ACKNOWLEDGEMENTS. We thank J. S. deBelle and D. Eydling for discussions and J. S. deBelle for language editing. The work was supported partly by the Deutsche Forschungsgemeinschaft and partly by a grant from the Human Frontiers Science Program to M.H.

working memory, and that this memory can be engaged to suppress as well as prescribe a response. These results establish, in a subset of prefrontal neurons, the iconic nature of the memory code, and suggest a role for visual memory in response suppression.

We examined the activity of neurons in the dorsolateral prefrontal cortex (Fig. 1a), in two rhesus monkeys trained on a compound oculomotor delayed-response task (Fig. 1b), in which on some trials, deferred saccades were directed to the location signalled by the visual cue (standard ODR task), whereas on other trials saccades were made in the opposite direction (anti-saccade oculomotor delayed response, AS-ODR). On anti-saccade trials, the monkeys learned to override the prepotent tendency to look toward the location of the remembered visual stimulus. Impairments on anti-saccade models have been demonstrated both in patients with large frontal lobe lesions³, and in patients suffering from schizophrenia⁴⁻⁶.

Among 108 prefrontal neurons recorded during both ODR and AS-ODR trials, 51 neurons had directionally specific delay period activity in the ODR task. We concentrate here on the delay period activity of these neurons. Of these 51 neurons, 44 have been histologically localized to the principal sulcus or immediately adjacent cortex. Visual, mnemonic and oculomotor activities of prefrontal neurons during standard ODR task performance, as well as surgical methods, recording procedures and data analysis techniques have been described previously⁷⁻⁹. Firing rates in all task-related responses were significantly greater than baseline rates (two-tailed unpaired Student's *t* statistic, alpha level 0.05).

FIG. 1 Schematic drawings of recording location and behavioural tasks. *a*, Region of the prefrontal cortex sampled during unit recording. *b*, The compound ODR/AS-ODR task. Each trial began with fixation of a central target (750 ms), either a spot (ODR trials) or a small plus sign (AS-ODR trials). A visual stimulus was then presented for 500 ms 13° to the left or to the right (location at random). Continual fixation was required during the presentation of the peripheral stimulus and throughout the 3-s delay period that followed. At the end of the delay, the fixation target disappeared, signalling the response phase. The monkey had 500 ms to direct a saccade in the dark to an invisible response window (6° diameter) centred either on the stimulus (ODR trials), or directly opposite the stimulus (AS-ODR trials). Correct responses were rewarded with a drop of juice. ODR and AS-ODR trials were either randomly interleaved within a session or given in blocks.

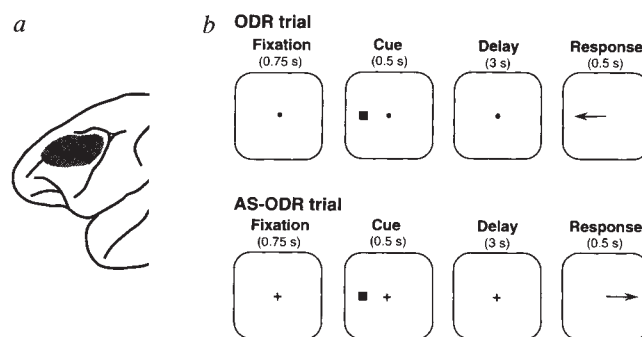
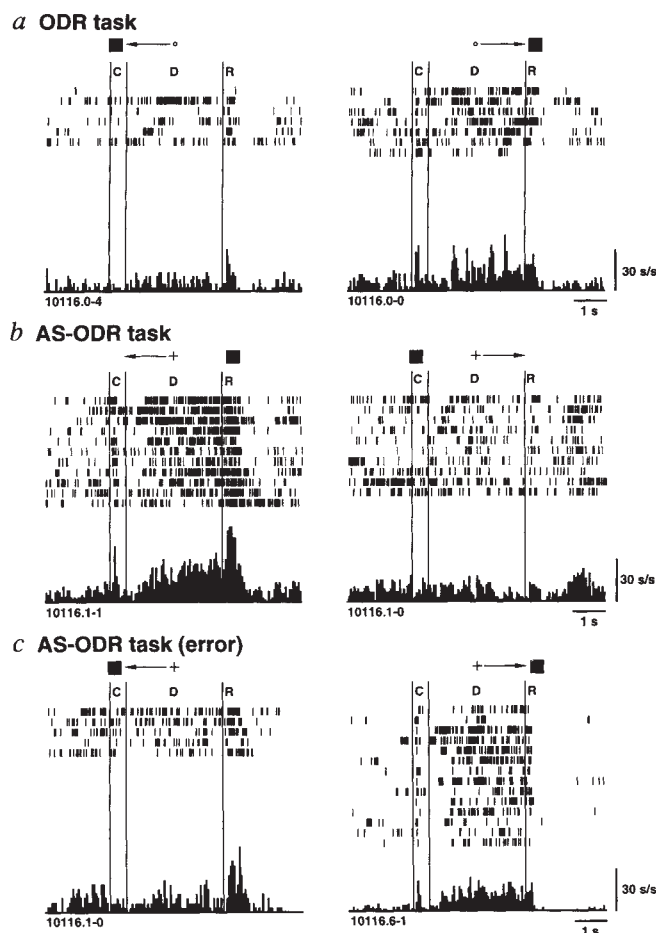


Figure 2 illustrates a principal sulcus neuron whose increased firing in the delay is linked to the location of the visual cue, not the direction of the saccade. First, the neuron responded 98 ms after the onset of the visual stimulus when it appeared on the right but not the left (Fig. 2*a*, right; *b*, left; *c*, right). Second, it showed increased activity on ODR trials during the 3-s delay period, again only after the rightward stimulus (Fig. 2*a*, right). Third, and most relevant to the memory code, on anti-saccade trials, activity was again elevated in the delay period following the rightward stimulus (Fig. 2*b*, left), even though the saccade at the end of the delay was directed to the left. Fourth, on AS-ODR error trials, when the monkey mistakenly directed its saccades toward the stimulus, neuronal activity in the delay again reflected the rightward location of the stimulus (Fig. 2*c*, right). Thus, the response of this neuron in the delay was keyed

to the direction of a preceding stimulus, and not to that of the impending saccade, whether that saccade was correct or incorrect, and whether it was toward or away from the original stimulus. A similar pattern of stimulus dependence was found in the delay period activity of 30 out of 51 (or 59%) of the prefrontal neurons studied. If the 44 histologically verified principal sulcus neurons are considered alone, this percentage rises to 68%. The preponderance of stimulus-dependent delay period activity in this area is strong evidence of a prefrontal specialization for ideational processing that does not rely on a motor code and is not mediated by motor signals.

Response-coding neurons were also found, in smaller numbers (25%, 13/51), intermixed among stimulus-coding neurons within the principal sulcus. An example of response-coding delay period activity in a principal sulcus neuron is shown in Fig. 3. Activity was

FIG. 2 Stimulus-dependent delay-period activity of a principal sulcus neuron (10116, left hemisphere). *a*, Delay period activity is elevated on ODR trials after rightward stimuli and before rightward saccades (right). *b*, Delay period activity is elevated on AS-ODR trials after the rightward stimulus, before leftward saccades (left). Because the response persists across trials with identical stimuli but opposite saccades, it codes the location of the stimulus. *c*, The response still occurs after rightward stimuli on AS-ODR error trials when the monkey mistakenly makes a saccade to the target (right). The neuron also responded phasically 98 ms after the onset of the visual stimulus on the right (cue (C) period: *a*, right; *b*, left; *c*, right), and 53 ms before saccades to the left (response (R) period: *a*, right; *b*, left; *c*, right). The independence of the preferred directions of visual and mnemonic activity on the one hand, and oculomotor activity on the other, is brought out by the inverted target-saccade relationship imposed by the anti-saccade task. In a two-way ANOVA (stimulus location by saccade direction), delay period activity following the rightward stimulus was significantly greater than that following the leftward stimulus ($F_{(\text{stimulus location})} = 19.56$, d.f. = 1, $P \leq 0.000$). Neither saccade direction ($F_{(\text{saccade direction})} = 0.78$, d.f. = 1, $P = 0.385$) nor the interaction term were significant ($F_{(\text{sl/sd})} = 2.70$, d.f. = 1, $P = 0.111$). The filled squares and arrows above each raster indicate the location of visual cue, and directions of saccadic eye movements, respectively. The abbreviations C, D, and R indicate cue, delay and response periods. Vertical lines through rasters indicate, left to right, the onset of the visual stimulus, the offset of the visual stimulus (beginning of the delay period), and the offset of the fixation spot (end of delay, beginning of response period). Vertical and horizontal bars at lower right corner of each pair of histograms indicate 30 spikes per s and 1 s, respectively.



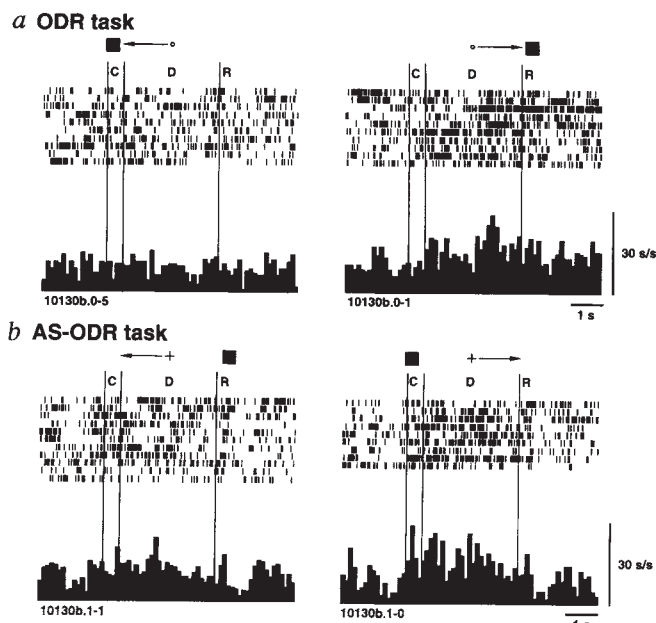


FIG. 3 Saccade-dependent delay-period activity of a principal sulcus neuron (10130b, right hemisphere). *a*, Delay period activity is elevated on ODR trials after rightward stimuli before rightward saccades (right). *b*, Delay period activity is elevated on anti-saccade trials after leftward stimuli before rightward saccades (right). Because the response persists across trials with identical saccades but opposite stimuli, it codes the direction of the saccade. In a two-way ANOVA (stimulus location by saccade direction) delay period activity preceding rightward saccades was significantly greater than that preceding leftward saccades ($F_{(\text{saccade direction})} = 12.44$, d.f. = 1, $P = 0.001$). Neither stimulus location ($F_{(\text{stimulus location})} = 1.74$, d.f. = 1, $P = 0.195$) nor the interaction term ($F_{(\text{sl, sd})} = 0.80$, d.f. = 1, $P = 0.376$) were significant. Conventions are the same as in Fig. 2.

greater in the delay before rightward saccades, irrespective of whether the stimulus signalling the saccade had appeared on the right (Fig. 3*a*, right) or the left (Fig. 3*b*, right). Similar motor-planning activity has been recorded from neurons in the frontal eye fields¹⁰, the posterior parietal cortex¹¹, the supplementary motor cortex and neostriatum¹², and in the superior colliculus¹³. The prefrontal region from which our recordings were made has connections to all of these structures^{14–20}, and could through these connections participate in the timing and/or direction of motor output. Eight of the 51 neurons (16%) failed to respond on AS-ODR trials and are not described further.

Evidence that stimulus-coding neurons may have a role in response suppression is provided by the neuron illustrated in Fig. 4, whose complex firing pattern is affected both by the location of the stimulus and the direction of the saccade. Delay period activity is increased before leftward saccades (Fig. 4*a*, left), but is suppressed after rightward stimuli, strongly on ODR trials (Fig. 4*a*, right), and transiently on AS-ODR trials (Fig. 4*b*, right). Given that the suppression is associated with the rightward location of the stimulus and not the rightward saccade, it may be driven by input from other stimulus coding prefrontal neurons similar to the one illustrated in Fig. 2. Independent response-coding and stimulus-coding mechanisms in prefrontal cortex might thus combine to mediate complex cognitive functions.

Stimulus-coding and response-coding delay period activities have been recorded previously from neurons in the principal sulcus on delayed-response tasks involving manual rather than oculomotor responses²¹. The agreement between those results and our own, obtained with tasks recruiting different motor sys-

tems, suggests that the concurrent representation of past stimuli and future actions is a general principle of prefrontal operation. This segregation may enable the flexible re-association of stimulus representations and motor plans on a moment to moment basis on tasks in which stimulus-response relationships are not fixed. Stimulus and response-coding prefrontal neurons may directly or indirectly contribute to activity changes observed in motor areas, such as the supplementary motor area, the motor cortex, and the putamen, where target-dependent and limb-dependent preparatory activity have also been recorded during a delayed-response task¹².

Working memory as characterized in studies of human cognition is more than a passive storage device but also a workspace for manipulation of symbolic representations²². Such a workspace may be needed to perform the anti-saccade task in monkeys, because this task requires a mental inversion analogous to

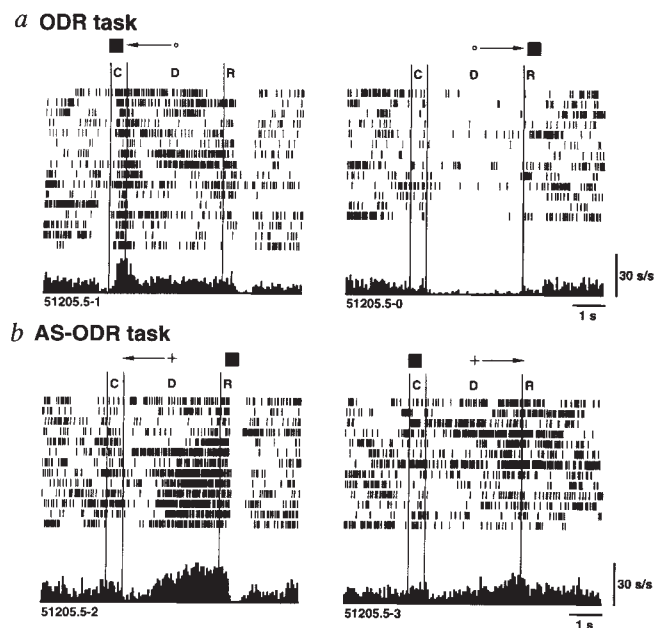


FIG. 4 Saccade-dependent delay period activity of a prefrontal neuron (51205, right hemisphere). *a*, Delay period activity on ODR trials is elevated before leftward saccades (left), and is strongly inhibited before rightward saccades (right). *b*, Delay period activity is again elevated before leftward anti-saccades (left), but inhibition before rightward anti-saccades is absent (right). Because the suppression seen on ODR trials did not occur before saccades of similar direction and amplitude on anti-saccade ODR trials, it cannot be attributed to the metrics of the impending eye movement. Conversely, a transient suppression is seen in the beginning of the delay on anti-saccade trials after the stimulus appears on the right. This transient suppression, and the more lasting suppression seen on ODR trials, have the rightward stimulus location in common. Thus the suppression is more closely associated with the rightward location of the stimulus than the rightward direction of the saccade. The suppression on anti-saccade trials following the rightward stimulus is overridden by an excitatory drive on this unit anticipating the leftward direction of the upcoming saccade. A visual response is evident after the stimulus appears on the left, if a saccade is made to the stimulus (*a*, left). The visual response is not present when the saccade is made away from the stimulus (*b*, right). This pattern of activity may reflect an enhancement of visual responsiveness to stimuli serving as saccade targets similar to that previously reported in the FEF¹⁰ and principal sulcus²³. A two-way ANOVA (stimulus location by saccade direction) confirms that delay period activity preceding leftward saccades was significantly greater than that preceding rightward saccades ($F_{(\text{saccade direction})} = 18.78$, d.f. = 1, $P = 0.000$). Stimulus location alone was not a significant factor ($F_{(\text{stimulus location})} = 0.64$, d.f. = 1, $P = 0.429$), but the interaction between stimulus location and saccade direction was significant ($F_{(\text{sl, sd})} = 19.79$, d.f. = 1, $P = 0.000$). The significant interaction term reveals that this neuron's pattern of activity in the delay is a function of both stimulus location and saccade direction.

processes engaged in humans performing an anti-saccade task³, the Stroop Test² or Wisconsin Card Sort Test¹. In these tasks, as in AS-ODR, habitual, generally sensory-driven, responses must be inhibited and less potent, generally instruction-guided, alternative responses selected. Loss of prefrontal neurons that hold instructional information 'on line' may explain why damage to prefrontal cortex commonly results not only in the absence of a correct response but distraction by sensory cues and disinhibition and preservation of competing habitual responses. As mentioned, schizophrenic patients are impaired on a variety of anti-saccade tasks⁴⁻⁶. If these deficits reflect the dysfunction of working memory mechanisms residing in prefrontal cortex, as seems reasonable, understanding the neural mechanism subserving delayed-response function in non-human primates could hold clues to cognitive dysfunction in mental illness as well. □

Received 30 March; accepted 9 August 1993.

1. Milner, B. *Arch. Neurol.* **9**, 90-100 (1963).

2. Perret, E. *Neuropsychologia* **12**, 323-330 (1974).
3. Guillon, D., Buchtel, H. A. & Douglas, R. M. *Expl Brain Res.* **58**, 455-472 (1985).
4. Park, S. & Holzman, P. *Arch. gen. Psychiat.* **49**, 975-982 (1993).
5. Fukushima, J. et al. *Biol. Psychiatry* **23**, 670-677 (1988).
6. Thaker, G. K. et al. *Psychopharmac. Bull.* **25**, 491-497 (1989).
7. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. *J. Neurophysiol.* **61**, 331-349 (1989).
8. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. *J. Neurophysiol.* **63**, 814-831 (1990).
9. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. *J. Neurophysiol.* **65**, 1464-1483 (1991).
10. Bruce, C. J. & Goldberg, M. E. *J. Neurophysiol.* **53**, 603-635 (1985).
11. Gnadt, J. W. & Andersen, R. A. *Expl Brain Res.* **70**, 216-220 (1988).
12. Alexander, G. E. & Crutcher, M. D. *J. Neurophysiol.* **64**, 164-178 (1990).
13. Mays, L. E. & Sparks, D. L. *J. Neurophysiol.* **43**, 207-232 (1980).
14. Barbas, H. & Mesulam, M. M. *J. comp. Neurol.* **200**, 407-431 (1981).
15. Cavada, C. & Goldman-Rakic, P. S. *J. comp. Neurol.* **287**, 422-445 (1989).
16. McGuire, P. K., Bates, J. F. & Goldman-Rakic, P. S. *Cereb. Cortex* **1**, 390-407 (1991).
17. Goldman, P. S. & Nauta, W. J. H. *Brain Res.* **116**, 145-149 (1976).
18. Fries, W. *J. comp. Neurol.* **230**, 55-76 (1984).
19. Barbas, H. & Pandya, D. N. *J. comp. Neurol.* **256**, 211-228 (1987).
20. Selemon, L. D. & Goldman-Rakic, P. S. *J. Neurosci.* **5**, 776-794 (1985).
21. Niki, H. & Watanabe, M. *Brain Res.* **105**, 79-88 (1976).
22. Baddeley, A. D. & Hitch, G. in *The Psychology of Learning and Motivation* Vol. 8 (ed. Bower, G. H.) 47-90 (Academic, New York, 1974).
23. Boch, R. A. & Goldberg, M. E. *J. Neurophysiol.* **61**, 1064-1084 (1989).

ACKNOWLEDGEMENTS. We thank C. J. Bruce for discussion in early stages of this investigation and F. A. W. Wilson, S. P. O. Scallidhe and M. Kritzer for suggestions on the manuscript and G. Leydon, M. Papero, S. Morgenstern, J. Coburn and M. Pappy for experimental assistance.

MIF is a pituitary-derived cytokine that potentiates lethal endotoxaemia

J. Bernhagen^{*†}, T. Calandra^{*}, R. A. Mitchell^{*}, S. B. Martin[‡], K. J. Tracey^{*§}, W. Voelker[†], K. R. Manogue^{*}, A. Cerami^{*} & R. Bucala^{*||}

^{*} The Picower Institute for Medical Research, 350 Community Drive, Manhasset, New York 11030, USA

[†] Physiologisch-chemisches Institut, University of Tübingen, W-7400 Tübingen, Germany

[‡] Division of Neurosurgery, New York Hospital-Cornell University Medical College, New York, New York 10021, USA

[§] Department of Surgery, North Shore University Hospital-Cornell University Medical College, Manhasset, New York 11030, USA

CYTOKINES are critical in the often fatal cascade of events that cause septic shock¹⁻³. One regulatory system that is likely to be important in controlling inflammatory responses is the neuroendocrine axis. The pituitary, for example, is ideally situated to integrate central and peripheral stimuli⁴, and initiates the increase in systemic glucocorticoids that accompanies host stress responses⁶⁻⁸. To assess further the contribution of the pituitary to systemic inflammatory processes, we examined the secretory profile of cultured pituitary cells and whole pituitaries *in vivo* after stimulation with bacterial lipopolysaccharide (LPS). Here we identify macrophage migration inhibitory factor (MIF)⁹⁻¹¹ as a major secreted protein released by anterior pituitary cells in response to LPS stimulation. Serum analysis of control, hypophysectomized and T-cell-deficient (nude) mice suggests that pituitary-derived MIF contributes to circulating MIF present in the post-acute phase of endotoxaemia. Recombinant murine MIF greatly enhances lethality when co-injected with LPS and anti-MIF antibody confers full protection against lethal endotoxaemia. We conclude that MIF plays a central role in the toxic response to endotoxaemia and possibly septic shock.

The anterior pituitary cell line AtT-20 was cultured serum-free in the presence of increasing amounts of bacterial endotoxin *Escherichia coli* 0111:B4 LPS). Conditioned media were analysed at different times for secreted proteins by SDS-polyacrylamide gel electrophoresis. Endotoxin stimulation resulted in a specific time- and concentration-dependent appearance of a

12.5K protein in the medium (Fig. 1a). This protein was isolated and its N-terminal sequence identified it as the murine homologue of human MIF⁹ (96% identity over 27 amino acids; Fig. 1b). Murine MIF was then cloned from complementary DNA prepared from stimulated pituitary cells, expressed in *E. coli* and used as immunogen for antibody production.

Under serum-free conditions, relatively high concentrations of LPS were necessary to demonstrate MIF release by silver staining. When analysed by western blotting with anti-MIF antibody, however, as little as 100 pg ml⁻¹ of LPS induced pituitary cell secretion of MIF (Fig. 1c). Supplementation of culture medium with 1% serum markedly increased secretion of MIF, consistent with studies indicating that LPS responses are potentiated by serum factors such as LPS-binding protein^{12,13}. In contrast, MIF was not released by incubation of pituitary cells with the inflammatory mediators tumour necrosis factor- α , interleukin-1 β , interleukin-6, or interferon- γ (data not shown). Western blotting also showed that resting, non-stimulated cells contained significant amounts of pre-formed MIF. Immunocytochemistry confirmed these results and showed that immunoreactive intracellular MIF is released within 16 h of LPS stimulation (Fig. 1d).

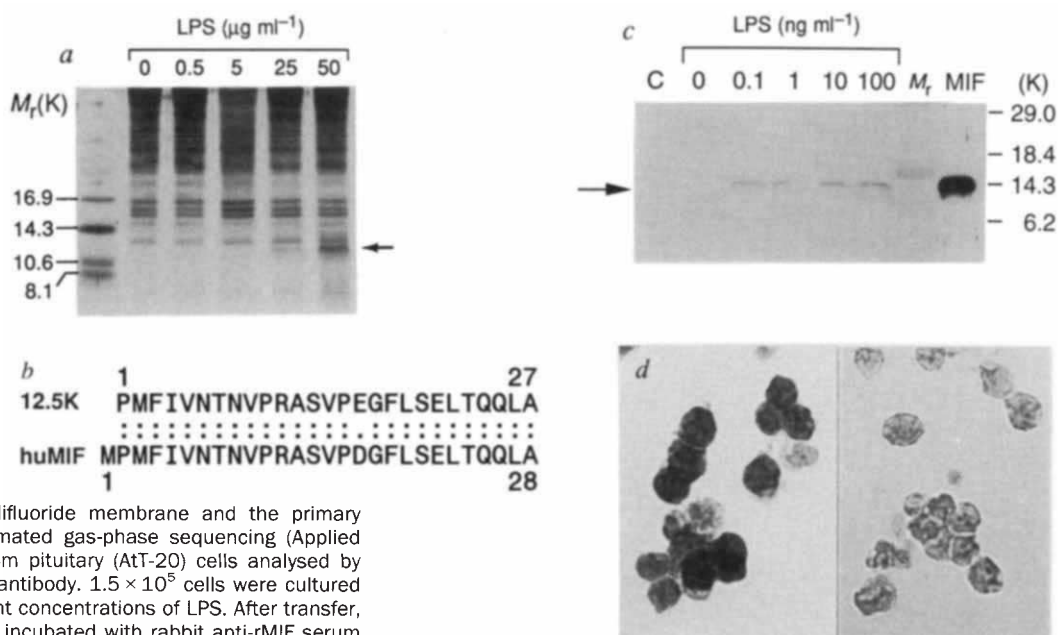
Human macrophage migration inhibition was one of the first cytokine activities to be described and results from product(s) elaborated by activated T cells^{10,11}. Cloned human MIF has an *M_r* of 12.5K and exerts a broad range of pro-inflammatory activities on monocytes/macrophages^{9,14,17}. We used reverse transcription-polymerase chain reaction (RT-PCR) analysis to investigate the expression of pituitary MIF *in vivo* in mice injected with sublethal amounts of LPS (2.25 mg kg⁻¹). Pituitary MIF messenger RNA levels increased with time and reached a plateau 16-24 h after LPS challenge (Fig. 2a). These findings were confirmed by competitive PCR analysis, which showed a 3-fold increase in pituitary MIF mRNA after LPS administration (Fig. 2b). Although infiltrating mononuclear cells were not evident in stained pituitary sections after LPS treatment, we wished to exclude a potential T-cell contribution of pituitary MIF mRNA. Pituitary cDNA analysis for the T-cell gene product CD2 was uniformly negative, ruling out infiltrating T-cells as a possible source of pituitary MIF mRNA (Fig. 2a). Also, LPS did not induce pituitary MIF mRNA in the genetically endotoxin-resistant mouse strain C3H/HeJ (Fig. 2c).

The pituitary content of MIF protein *in vivo* was analysed by western blotting of pituitary lysates. As predicted from studies of cultured pituitary cells, pituitaries from normal, non-stimulated mice showed substantial amounts of pre-formed MIF protein (Fig. 3a). There was a significant decrease in pituitary MIF con-

|| To whom correspondence should be addressed.

FIG. 1 Identification of murine MIF in pituitary (AtT-20) cell supernatants. *a*, SDS-PAGE (18%) with silver staining of conditioned media from cells cultured for 24 h in enriched serum-free Joklik's medium with increasing concentrations of LPS. Medium (200 μ l) was removed from 1.5×10^5 cells and concentrated 10-fold by membrane filtrations (10K cut-off). After LPS stimulation, the 12.5K protein band increased in intensity from 8 to 48 h (not shown). *b*, N-terminal amino-acid sequence homology between the LPS-induced 12.5K protein and human MIF. The 12.5K protein was transferred to polyvinylidene difluoride membrane and the primary sequence determined by automated gas-phase sequencing (Applied Biosystems). *c*, MIF release from pituitary (AtT-20) cells analysed by western blotting with anti-rMIF antibody. 1.5×10^5 cells were cultured serum-free for 18 h with different concentrations of LPS. After transfer, nitrocellulose membranes were incubated with rabbit anti-rMIF serum (1:1,000) and antibody visualized with horseradish peroxidase-conjugated goat anti-rabbit antibody. rMIF (100 ng) was electrophoresed on the same gel. *d*, Immunocytochemical analysis of MIF secretion from cultured pituitary (AtT-20) cells. Cells were cultured for 16 h in the presence (right panel) or absence (left panel) of $25 \mu\text{g ml}^{-1}$ LPS, washed, fixed with formaldehyde and incubated with anti-rMIF serum (1:1,000 dilution) before staining with immunoperoxidase-linked goat anti-rabbit antibody²².

METHODS. Murine MIF was cloned by DNA amplification of cDNA prepared from LPS-stimulated AtT-20 cells. Nucleotide sequencing of the



subcloned cDNA showed high homology to human MIF⁹ (88.2% identity over 348 nucleotides and 88.7% identity over 115 amino acids). The AtT-20 MIF cDNA sequence is entered in the EMBL Data Bank under accession number Z23048. Murine MIF cDNA was subcloned into the *E. coli* expression vector pET11b (Novagen) and the expressed protein purified by Fast Protein Liquid Chromatography using Mono-Q resin (Pharmacia). Bioactivity was established by monocyte migration inhibition in modified Boyden chambers²³. Rabbits were immunized with rMIF and antibody production assessed by western blotting.

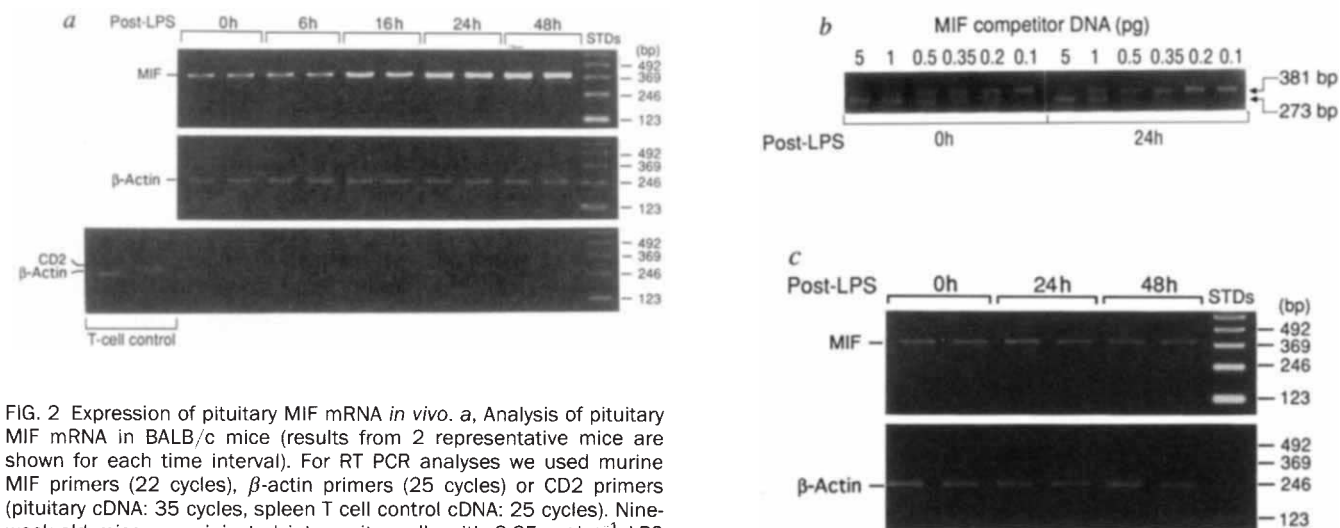


FIG. 2 Expression of pituitary MIF mRNA *in vivo*. *a*, Analysis of pituitary MIF mRNA in BALB/c mice (results from 2 representative mice are shown for each time interval). For RT-PCR analyses we used murine MIF primers (22 cycles), β -actin primers (25 cycles) or CD2 primers (pituitary cDNA: 35 cycles, spleen T cell control cDNA: 25 cycles). Nine-week-old mice were injected intraperitoneally with 2.25 mg kg^{-1} LPS and killed at the intervals shown. RNA was isolated with RNAzol (Biotecx Lab.) from a total of 4 mice per time interval. After reverse transcription with oligo(dT), DNA amplification products were analysed by agarose gel electrophoresis. Reverse transcription efficiency was normalized relative to β -actin mRNA. The murine MIF primers: 5'-CCATGCC-TATGTCATCGTG-3' and 3'-GAACAGCGGTGCAGGTAAGTG-5' amplify a 381-base-pair sequence that spans at least one intron (determined by Southern hybridization; unpublished results). The murine intron-spanning CD2 primers were 5'-CCTGGTCGACAGAGTTAA-3' and 3'-TCTGTTCTTGCAGACC-5'; β -actin primers were 5'-GTGGGCCGCTCT-AGGCACCA-3' and 3'-TGGCCTTAGGGTTCAGGGGG-5' (ref. 24). *b*, Competitive PCR quantification of pituitary MIF mRNA from BALB/c mice. Pituitary mRNA was analysed after LPS injection (2.25 mg kg^{-1}) by competitive PCR^{25,26} using a 273-bp competitive MIF template. The

competitive template was added in the range of 0.1–5 pg and co-amplified with cDNA using the MIF primers described in *a*. At 0 h, equivalent amounts of cDNA and competitor DNA amplification products appear in the lane showing 0.35 pg of MIF competitor template. At 24 h, equivalent amounts of cDNA and competitor DNA amplification products appear in the lane showing 1.0 pg of MIF competitor template. *c*, Expression of pituitary MIF mRNA in endotoxin-resistant C3H/HeJ mice (analyses from 2 representative mice are shown for each time interval). Nine-week-old mice ($n=4$ per time point) were injected intraperitoneally with LPS (2.25 mg kg^{-1}) and killed at the times shown. Congenic endotoxin-sensitive C3H/HeN mice injected with LPS showed an increase in MIF mRNA between 0 and 24 h that was similar to BALB/c mice (*a*).

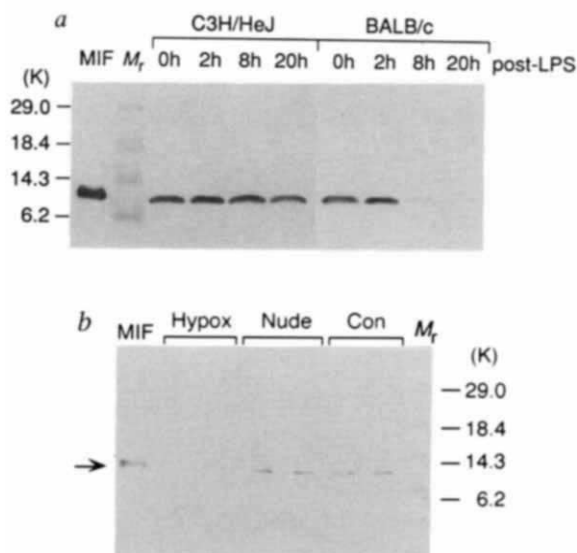


FIG. 3 Expression of MIF protein *in vivo*. **a**, Pituitary MIF protein content in C3H/HeJ and BALB/c mice challenged with LPS. A similar LPS-dependent decrease in pituitary MIF content was observed both in C3H/HeJ and BALB/c^{nu/nu} mice as in BALB/c mice (not shown). Mice ($n=2-5$ per time point) were injected intraperitoneally with LPS (2.25 mg kg^{-1}), pituitary lysates were prepared at intervals shown, and protein analysed by western blotting. Aliquots representing one quarter of a pituitary were electrophoresed, transferred to nitrocellulose membrane and incubated with anti-rMIF serum (1:1,000 dilution). Bound antibody was visualized with horseradish peroxidase-conjugated goat anti-rabbit antibody. Recombinant murine MIF (50 ng) was electrophoresed in the MIF standard lane. Single bands corresponding to MIF were quantified by laser densitometry (Shimadzu FDU-3, CS-9,000 U). Recombinant MIF standards were electrophoresed and transferred in adjacent lanes. Scanning integrals were calculated as relative MIF content normalized to the respective 0 h zone on each membrane. **b**, Western blotting analysis of serum for MIF protein. Serum was obtained from two representative hypophysectomized BALB/c (Hypox), T-cell-deficient BALB/c^{nu/nu} (Nude), and control BALB/c (Con) mice. Hypophysectomized mice were obtained from Charles River (Kingston, NY) and maintained with drinking water supplemented with 5% glucose. Mice (9 weeks old) were injected intraperitoneally with LPS at 2.25 mg kg^{-1} and serum samples (5 μl each) analysed by western blotting with anti-MIF antibody. Hypophysectomized animals were necropsied to ensure complete removal of pituitary. Blots were developed as described for **a**. As expected, serum MIF was not detected after administration of LPS to intact, endotoxin-resistant C3H/HeJ mice (data not shown).

tent 8–20 h after LPS administration in endotoxin-sensitive mice. Using densitometric scanning (Fig. 3a), we estimated the content of stored, pre-formed MIF within a single pituitary to be about 40 ng, corresponding to 0.05% of total pituitary protein. By comparison, the pituitary hormones ACTH and prolactin comprise 0.2 and 0.08% respectively, of total (human) pituitary protein¹⁸.

We then assessed the contribution of pituitary MIF to MIF that might appear in serum during endotoxaemia. Control (BALB/c) and T-cell-deficient (BALB/c^{nu/nu}) mice were injected with a sublethal amount of LPS and serum samples withdrawn and analysed for MIF by western blotting. Serum MIF was undetectable at 0 h and increased to maximum levels 20 h after LPS stimulation (not shown). This time course is consistent with the disappearance of pituitary MIF protein and the induction of pituitary MIF mRNA levels *in vivo* (Figs 2a, 3a). In contrast, specimens from LPS-stimulated hypophysectomized BALB/c mice had no detectable MIF in serum at 20 h, the time at which serum MIF levels were highest in control and T-cell-deficient mice (Fig. 3b). These data indicate that the pituitary is an important source of the MIF that appears in serum after LPS adminis-

tration. Nevertheless, we cannot exclude the possibility that in addition, pituitary mediators induce the release of T-cell or non-T-cell sources of MIF that account for the appearance of MIF in serum during endotoxaemia.

The observation that hypophysectomy sensitizes animals to endotoxin lethality¹⁹ suggests that pituitary MIF might improve host resistance to lethal endotoxaemia. To investigate the role of MIF in experimentally induced septic shock, we injected recombinant murine MIF (rMIF) together with LPS into BALB/c mice. Rather than conferring protection, rMIF markedly potentiated LPS lethality (85% in the rMIF + LPS group versus 35% in the LPS group; Fig. 4a). Administration of rMIF by itself produced the visible symptoms of endotoxaemia, but without lethality in the dose range studied ($10-50 \text{ mg kg}^{-1}$). To confirm the participation of MIF in LPS-induced lethality, we examined the effect of anti-rMIF serum in mice that had been given an otherwise lethal dose of LPS. Anti-rMIF serum fully protected mice against endotoxaemia, increasing survival from 50 to 100% (Fig. 4b).

These findings indicate that MIF is important in endotoxic shock. Although the function of MIF in the host response to tissue invasion remains to be explained, pituitary production of MIF may reflect the interplay of diverse neurohumoral stimuli and serve to regulate systemic inflammatory responses. The

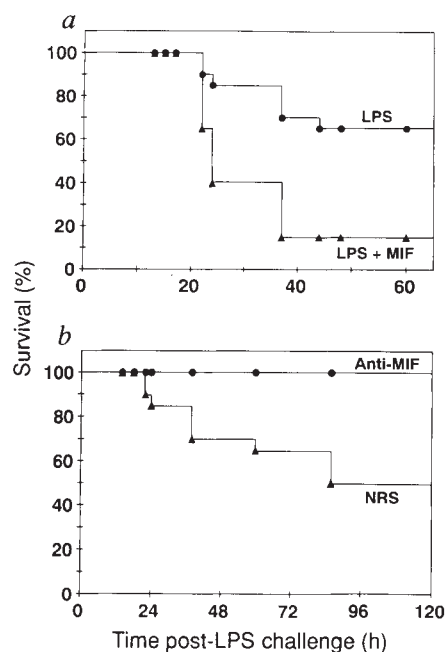


FIG. 4 Role of MIF in endotoxic shock. **a**, MIF potentiates lethal endotoxaemia in mice. Nine-week-old BALB/c mice (19–21 g) were injected intraperitoneally with LPS at a dose of 15 mg kg^{-1} and rMIF at 5 mg kg^{-1} (200 μl , containing $<45 \text{ pg}$ of LPS per μg rMIF) or with LPS at 15 mg kg^{-1} in saline (200 μl). Mice were given a second injection of 5 mg kg^{-1} rMIF (●), or saline (▲), 12 h after the initial challenge. Saline alone was adjusted to contain LPS in an amount equivalent to that determined to be present in rMIF (50 ng ml^{-1}). Data points are from two independent experiments (10 mice per group; $P=0.003$ by two-tailed Fisher's exact test. **b**, Anti-MIF antibody protects against lethal endotoxaemia in mice. Nine-week-old BALB/c mice (19–21 g) were injected intraperitoneally with 200 μl anti-rMIF serum (containing $<0.2 \text{ ng LPS ml}^{-1}$) or normal rabbit serum (NRS, containing $<3 \text{ ng LPS ml}^{-1}$) 2 h before intraperitoneal LPS administration (17.5 mg kg^{-1} , an LD₇₅ for saline-injected control mice established in previous dose calibration experiments). Data points are from two independent experiments, each group consisting of 10 mice; $P=0.0004$ by the two-tailed Fisher's exact test. We established that anti-rMIF serum and NRS did not crossreact with *E. coli* 0111:B4 LPS by western blotting and LPS-specific ELISA²⁷.

increased sensitivity to endotoxin that has been reported after hypophysectomy¹⁹ suggests that MIF functions together with other pituitary factors such as ACTH to modulate both pro- and anti-inflammatory effects. Of note, cytokines such as TNF- α exert both beneficial and detrimental effects on the host depending on the tissue site, the timing and the amount of cytokine produced^{20,21}.

Our *in vitro* experiments indicate that anterior pituitary cells specifically release MIF in response to LPS. Nevertheless, we cannot exclude a mechanism for cytokine-mediated release of pituitary MIF *in vivo*. Circulating cytokines may act together with central, neurally derived stimuli to induce pituitary MIF release. Studies are in progress to determine the precise neuro-humoral stimuli and feedback loops that control pituitary MIF secretion. Investigation of these pathways may lead to new strategies for the treatment of septic shock and other inflammatory conditions. □

Received 19 July; accepted 18 August 1993.

1. Beutler, B. & Cerami, A. *Nature* **320**, 584–588 (1986).
2. Tracey, K. J. & Lowry, S. F. *Adv. Surgery* **23**, 21–56 (1990).
3. Dinarello, C. A. *J. infect. Dis.* **163**, 1177–1184 (1991).
4. Koenig, J. I. et al. *Endocrinology* **126**, 3053–3057 (1990).
5. Besedovsky, H., Del Rey, A., Sorkin, E. & Dinarello, C. A. *Science* **233**, 652–654 (1986).
6. Reichlin, S. *Williams Textbook of Endocrinology* 8th edn. 135–219 (Saunders, Philadelphia, 1992).
7. Yasuda, N. & Greer, M. A. *Endocrinology* **102**, 947–953 (1978).

8. Makara, G. B., Stark, E. & Meszaros, T. *Endocrinology* **88**, 412–417 (1971).
9. Weiser, W. Y. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 7522–7526 (1989).
10. Bloom, B. R. & Bennett, B. *Science* **153**, 80–82 (1966).
11. David, J. R. *Proc. natn. Acad. Sci. U.S.A.* **65**, 72–77 (1966).
12. Schumann, R. R. et al. *Science* **249**, 1429–1431 (1990).
13. Wright, S. D., Ramos, R. A., Tobias, P. S., Ulevich, R. J. & Mathison, J. C. *Science* **249**, 1431–1433 (1990).
14. Weiser, W. Y., Pozzi, L. M. & David, J. R. *J. Immun.* **147**, 2006–2011 (1991).
15. Pozzi, L. M. & Weiser, W. Y. *Cell. Immun.* **145**, 372–379 (1992).
16. Weiser, W. Y., Pozzi, L. M., Titus, R. G. & David, J. R. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8049–8052 (1992).
17. Cunha, F. Q. et al. *J. Immun.* **150**, 1908–1912 (1993).
18. Thorner, M. O., Vance, M. L., Horvath, E. & Kovacs, K. *Williams Textbook of Endocrinology* 8th edn 221–310 (Saunders, Philadelphia, 1992).
19. Silverstein, R., Turley, B. R., Christoffersen, C. A., Johnson, D. C. & Morrison, D. C. *J. exp. Med.* **173**, 357–365 (1991).
20. Tracey, K. J., Vlassara, H. & Cerami, A. *Lancet* **i**, 1122–1126 (1989).
21. Fraker, D. L., Stovroff, M. C., Merino, M. J. & Norton, J. A. *J. exp. Med.* **168**, 95–105 (1988).
22. Kriegler, M. *Gene Transfer and Expression: A Laboratory Manual* 219–221 (Freeman, New York, 1990).
23. Boyden, S. V. *J. exp. Med.* **115**, 453–466 (1962).
24. Alonso, S., Minty, A., Bourlet, Y. & Buckingham, M. J. *molec. Evol.* **23**, 11–22 (1986).
25. Gilliland, G., Perrin, S., Blanchard, K. & Bunn, H. F. *Proc. natn. Acad. Sci. U.S.A.* **87**, 2725–2729 (1990).
26. Li, B. et al. *J. exp. Med.* **174**, 1259–1262 (1991).
27. Heumann, D., Baumgartner, J. D., Jacot-Guillarmod, H. & Glauser, M. P. *J. infect. Dis.* **163**, 762–768 (1991).

ACKNOWLEDGEMENTS. We thank J. Fernandez and the Rockefeller University biopolymer facility for protein sequencing; B. Sherry for discussions and for assistance with the migration assay; M. Knaup and D. Heumann for the LPS-ELISA; B. Manogue for assistance with pituitary sectioning; and J. R. David for his unpublished murine MIF cDNA sequence. This work was supported by the NIH, the Arthritis Foundation, by a grant from the Fonds de Perfectionnement du Centre Hospitalier Universitaire Vaudois, and the Swiss Foundation for Research in Medicine and Biology.

Pathophysiological role of endothelin revealed by the first orally active endothelin receptor antagonist

Martine Clozel, Volker Breu, Kaspar Burri, Jean-Marie Cassal, Walter Fischli, Gillian A. Gray, Georges Hirth, Bernd-Michael Löffler, Marcel Müller, Werner Neidhart & Henri Ramuz

Pharma Division, Preclinical Research, F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, CH-4002 Basel, Switzerland

SINCE its discovery¹, endothelin-1 has attracted considerable scientific interest because of its extremely potent and long-lasting vasoconstrictor effect and its binding to G-protein-coupled receptors². Plasma concentrations of endothelin-1 are low³ and its release by endothelial cells is polarized towards the basolateral side^{4,5}, suggesting that it is a paracrine factor and not a hormone. Consequently, the effect of injected endothelin-1 may not reflect the effect of endogenous endothelin-1. In contrast, blockade of the action of endogenous endothelin-1 using receptor antagonists should be a valuable means of investigating its physiological and pathological effects. We report here evidence for the pathophysiological role of endothelin-1 as brought by the first synthetic orally active non-peptide antagonist of endothelin receptors, Ro 46-2005.

Our goal was to develop mixed antagonists of all known endothelin (ET) receptors and to use them to assess the potential role of ET-1 in models of focal vasoconstriction. Two subtypes of ET receptors, ET_A and ET_B, have been cloned and characterized^{6,7}, although there is evidence for other subtypes^{8,9}. The ET_A receptor mediates a major part of the vasoconstriction induced by ET-1 (ref. 10). The ET_B subtype mediates endothelium-dependent vasodilation¹¹, but its stimulation by selective agonists can also induce vasoconstriction^{12–15}. In addition, because ET was suspected to play a role in chronic diseases^{16–18}, the development

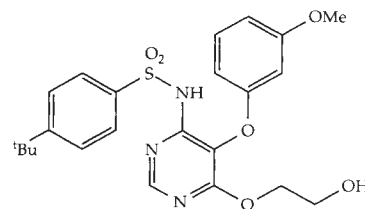


FIG. 1 Structure of Ro 46-2005.

of non-peptidic, small M_r antagonists that could be orally administered was an important objective. For this purpose several thousands of compounds from a chemical library were screened for their capacity to inhibit specific ¹²⁵I-ET-1 binding in a human placenta membrane preparation¹⁹. This screening led to the identification of a class of pyrimidinyl sulphonamides that had been synthesized as part of an antidiabetic project and were noted to be weak inhibitors of ¹²⁵I-ET-1 binding. Structural modification of these compounds by chemical synthesis led to the discovery of Ro 46-2005 (4-tert-butyl-N-[6-(2-hydroxyethoxy)-5-(3-methoxy-phenoxy)-4-pyrimidinyl]-benzenesulphonamide), a potent inhibitor itself devoid of hypoglycaemic activity. Its structure is shown in Fig. 1.

Before testing Ro 46-2005 *in vivo*, it was crucial to assess its binding and functional characteristics, and its selectivity. Ro 46-2005 completely inhibited the specific binding of ¹²⁵I-ET-1 to human vascular smooth muscle cells (ET_A receptor) and rat aortic endothelial cells (ET_B receptor), with half-maximal inhibition at concentrations of 2.2×10^{-7} M and 1×10^{-6} M, respectively. The *in vitro* functional activity of Ro 46-2005 was examined on vascular preparations carrying ET_A and ET_B receptors. In isolated rat aortic rings denuded of their endothelium (ET_A receptors), Ro 46-2005 had no agonistic effect, but antagonized the contractions induced by ET-1 in a concentration-dependent manner (Fig. 2a). Increasing concentrations of Ro 46-2005 caused a parallel shift to the right of the concentration-response curves of ET-1, without causing a decrease in the maximal response. Analysis of the data²⁰ yielded

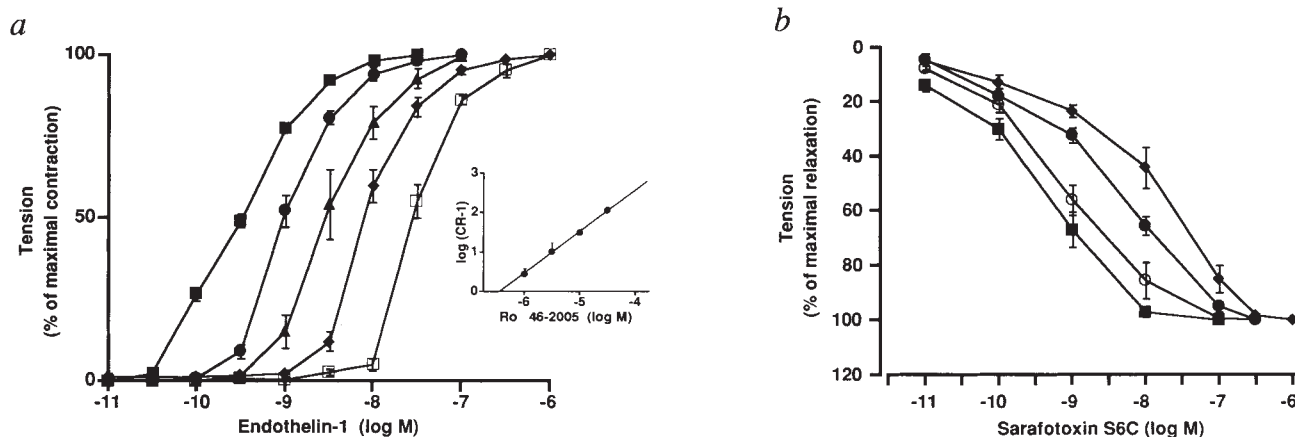


FIG. 2 *In vitro* antagonism of ET_A and ET_B receptors by Ro 46-2005. **a**, Effect of Ro 46-2005 (■, control vehicle; ●, 10^{-6} M; ▲, 3×10^{-6} M; ◆, 10^{-5} M; □, 3×10^{-5} M) on the constrictor effect of ET-1 in isolated rat aortic rings. Each point represents the mean $\pm$ s.e.m. of 4–7 experiments. In some cases the error bars are hidden within the symbols. Aortic rings were isolated from anaesthetized male Wistar-Kyoto rats²⁸, their endothelium was rubbed and the rings were suspended in organ baths filled with oxygenated Krebs-Henseleit solution under a tension of 3 g. The absence of functional endothelium was confirmed by absence of relaxation to acetylcholine (10μ M). After 10 min incubation with various concentrations of Ro 46-2005, cumulative doses of ET-1 were added. Inset: Schild plot of $\log$ [concentration ratio (CR)-1] as a function of antagonist concentration. A linear regression analysis using individual $\log$ (CR-1) values yielded a correlation coefficient of 0.95 and a slope of 1.05 ± 0.08 . **b**, Effect of Ro 46-2005 (■, control vehicle; ○,

10^{-7} M; ●, 10^{-6} M; ◆, 10^{-5} M) on endothelium-dependent relaxation to sarafotoxin S6c in isolated rat small mesenteric arteries. Each point represents the mean $\pm$ s.e.m. of 5–8 experiments. Segments from the third branch of the superior mesenteric artery (100 μ m internal diameter) were isolated from male Wistar rats and mounted in an isometric myograph under a resting tension of 200 mg (ref. 29). The presence of endothelium was confirmed by relaxation of the vessels to acetylcholine (10μ M) after precontraction with noradrenaline (5μ M). After washing and equilibration, the vessels were submaximally constricted by methoxamine (5μ M) in presence of 10 mM potassium chloride and reached a tension of 1.4 ± 0.1 g. After 10 min incubation with various concentrations of Ro 46-2005, cumulative doses of sarafotoxin S6c were added. Sarafotoxin S6c elicited a dose-dependent relaxation with a maximal relaxation of $64 \pm 6\%$ of the tension induced by methoxamine.

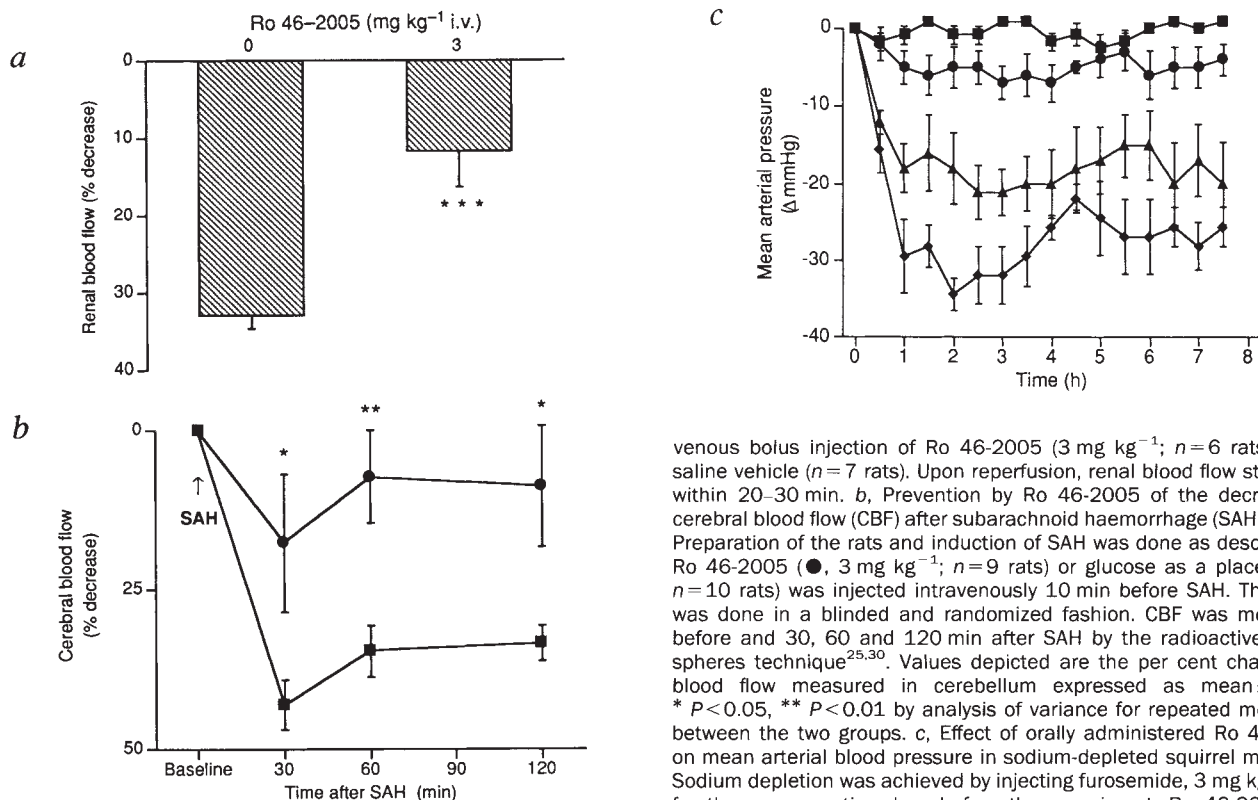


FIG. 3 *In vivo* effects of Ro 46-2005 in three different animal models. **a**, Prevention by Ro 46-2005 of post-ischaemic renal vasoconstriction in rats. Values depicted are per cent decreases in renal blood flow expressed as mean $\pm$ s.e.m. *** $P < 0.001$ by unpaired Student's *t*-test. Wistar rats were anaesthetized and instruments attached as described²². Renal ischaemia was induced by a 45-min clamping of the left renal artery via a snare placed around the artery at its origin. Ten min before induction of renal ischaemia, the rats received an intra-

venous bolus injection of Ro 46-2005 (3 mg kg^{-1} ; $n=6$ rats) or its saline vehicle ($n=7$ rats). Upon reperfusion, renal blood flow stabilized within 20–30 min. **b**, Prevention by Ro 46-2005 of the decrease in cerebral blood flow (CBF) after subarachnoid haemorrhage (SAH) in rats. Preparation of the rats and induction of SAH was done as described²⁵. Ro 46-2005 (●, 3 mg kg^{-1} ; $n=9$ rats) or glucose as a placebo (■, $n=10$ rats) was injected intravenously 10 min before SAH. The study was done in a blinded and randomized fashion. CBF was measured before and 30, 60 and 120 min after SAH by the radioactive micro-spheres technique^{25,30}. Values depicted are the per cent changes in blood flow measured in cerebellum expressed as mean $\pm$ s.e.m. * $P < 0.05$, ** $P < 0.01$ by analysis of variance for repeated measures between the two groups. **c**, Effect of orally administered Ro 46-2005 on mean arterial blood pressure in sodium-depleted squirrel monkeys. Sodium depletion was achieved by injecting furosemide, 3 mg kg^{-1} i.m. for three consecutive days before the experiment. Ro 46-2005 (■, control vehicle; $n=6$; ●, 10 mg kg^{-1} ; $n=5$; ▲, 30 mg kg^{-1} ; $n=5$; ◆, 100 mg kg^{-1} ; $n=4$) was administered by gavage. Blood pressure was monitored in conscious unrestrained monkeys chronically instrumented with a telemetry system³¹. Before drug administration (time 0) the arterial pressure was allowed to stabilize for at least 30 min. Baseline mean arterial blood pressure was 96 to 99 mmHg. Values depicted represent change in mean arterial blood pressure expressed as mean $\pm$ s.e.m. In some cases the error bars are hidden within the symbols.

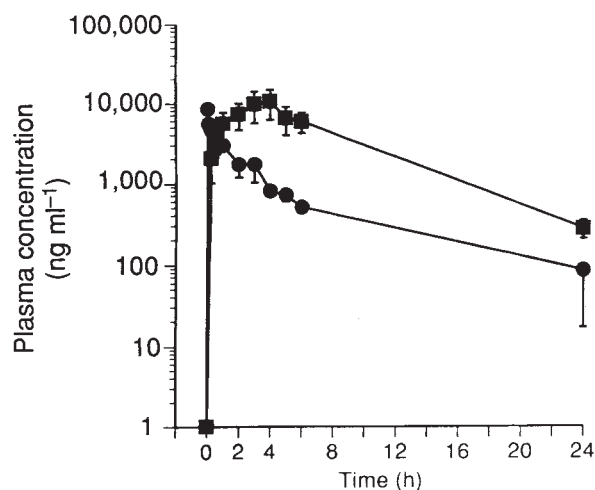


FIG. 4 Plasma concentrations of Ro 46-2005 after oral (■, 10 mg kg^{-1}) or intravenous (●, 1 mg kg^{-1}) administration in rats, as measured by a binding assay. Each point represents the mean $\pm$ s.e.m. of 4 experiments. Wistar rats were fitted with arterial and venous catheters under anaesthesia. They were allowed to recover for several hours and Ro 46-2005 was then administered intravenously or by oral gavage. Blood samples were withdrawn through the arterial catheter in the unrestrained rats and centrifuged. Ro 46-2005 was extracted from plasma using extraction by methanol. Ro 46-2005 was measured by its competition with ^{125}I -endothelin-1 for binding on recombinant human ET_A receptor expressed in a baculovirus/insect cell system. Plasma concentration of Ro 46-2005 was computed from a calibration curve of Ro 46-2005 in spiked rat plasma.

a pA_2 value (negative logarithm of the molar concentration of antagonist which causes a twofold parallel shift to the right of the concentration-response curve) of 6.50 ± 0.04 and a slope of 1.05 ± 0.08 , indicating that Ro 46-2005 behaves as a purely competitive antagonist at the ET_A receptor. In isolated rat small mesenteric arteries precontracted by methoxamine, Ro 46-2005 inhibited endothelium-dependent relaxation induced by the selective ET_B receptor agonist sarafotoxin S6c (ref. 21) in a concentration-dependent manner (Fig. 2b). The shift to the right of the concentration-response curve of sarafotoxin S6c was parallel, with no significant decrease in the maximal relaxation, indicating a competitive antagonism at the ET_B receptor as well. The pA_2 value of Ro 46-2005 on ET_B receptor was 6.47 ± 0.16 .

Knowing the almost irreversible binding of ET-1 to its receptors, it was remarkable to see that Ro 46-2005 was able not only to prevent the contractions but also to reverse an established contraction to ET-1, with a half-maximal effect at a concentration of $7.8 \times 10^{-7} \text{ M}$. Finally, Ro 46-2005 is extremely selective for ET because, up to a concentration of $3 \times 10^{-5} \text{ M}$, a dose comparable to the maximal plasma concentrations reached after *in vivo* administration of pharmacologically active doses, it did not affect the contractions evoked by angiotensin II, serotonin, noradrenaline, potassium chloride or prostaglandin $\text{F}_{2\alpha}$. The functional antagonism of ET-1 by Ro 46-2005 and its specificity were both confirmed through *in vivo* experiments.

The evidence for a major role of ET-1 in the mediation of pathological vasospasm was shown by the efficacy of Ro 46-2005 in two models of vasoconstriction. In one model, anaesthetized rats subjected to a 45-min clamping of the left renal artery developed prolonged renal vasoconstriction on reperfusion. We have shown selective alteration in binding and functional

response to endothelin in this model of renal ischaemia²², and the efficacy of the ET_A receptor peptide antagonist BQ-123 has been proved in a comparable model of ischaemia-induced renal vasoconstriction²³. But BQ-123 also inhibits the action of angiotensin II (ref. 24) and may be less specific than Ro 46-2005. Ro 46-2005 (3 mg kg^{-1} i.v.) significantly reduced renal vasoconstriction during reperfusion (Fig. 3a). In a second model, anaesthetized rats were subjected to subarachnoid haemorrhage by injection of autologous blood in the cisterna magna. ET has been suggested to play a role in this animal model of cerebral vasoconstriction^{25,26} and in late cerebral vasospasm following subarachnoid haemorrhage in man²⁷. Cerebral blood flow decreased by about 40% 30 to 120 min after subarachnoid haemorrhage. Ro 46-2005 (3 mg kg^{-1} i.v.) dramatically reduced this cerebral vasoconstriction (Fig. 3b). In these two studies, Ro 46-2005 had no significant effect on systemic arterial blood pressure. In contrast, in sodium-depleted squirrel monkeys, orally administered Ro 46-2005 ($10\text{--}100 \text{ mg kg}^{-1}$) markedly decreased arterial blood pressure (Fig. 3c), without altering heart rate. This indicates that Ro 46-2005 is orally active. The results of serial measurements of binding inhibitory activity in plasma after oral and intravenous injection in rats showed that Ro 46-2005 has an oral bioavailability of about 30% (Fig. 4).

The efficacy of Ro 46-2005 in different experimental models shows the pathophysiological role of ET-1 and the potential of this new therapeutic concept of ET receptor blockade in the treatment of vasoconstriction. The validity of this concept in man remains to be determined, but the availability of orally active compounds should now make it possible to test this hypothesis not only in acute situations (cerebral vasospasm or renal ischaemia), but also in chronic diseases such as atherosclerosis, congestive heart failure and hypertension. □

Received 8 April; accepted 6 August 1993.

- Yanagisawa, M. et al. *Nature* **332**, 411–415 (1988).
- Kasuya, Y., Takuwa, Y., Yanagisawa, M., Masaki, T. & Goto, K. *Br. J. Pharmac.* **107**, 456–462 (1992).
- Ando, K., Hirata, Y., Shichiri, M., Emori, T. & Marumo, F. *FEBS Lett.* **245**, 164–166 (1989).
- Wagner, F. et al. *J. Biol. Chem.* **267**, 16066–16068 (1992).
- Yoshimoto, S. et al. *J. Cardiovasc. Pharmac.* **17**(suppl. 7), S260–S263 (1991).
- Arai, H., Hori, S., Aramori, I., Ohkubo, H. & Nakanishi, S. *Nature* **348**, 730–732 (1990).
- Sakurai, T. et al. *Nature* **348**, 732–735 (1990).
- Emori, T., Hirata, Y. & Marumo, F. *FEBS Lett.* **263**, 261–264 (1990).
- Sokolovsky, M., Ambar, I. & Gailron, R. *J. Biol. Chem.* **267**, 20551–20554 (1992).
- Ihara, M. et al. *Life Sci.* **50**, 247–255 (1993).
- Takayanagi, R. et al. *FEBS Lett.* **282**, 103–106 (1991).
- Clozel, M., Gray, G. A. & Breu, V. *Biochem. biophys. Res. Commun.* **186**, 867–873 (1992).
- Moreland, S., McMullen, D. M., Delaney, C. L., Lee, V. G. & Hunt, J. T. *Biochem. biophys. Res. Commun.* **184**, 100–106 (1992).
- McMurdo, L., Corder, R., Thiernemann, C. & Vane, J. R. *Br. J. Pharmac.* **108**, 557–561 (1993).

- Harrison, V. J., Randrianosa, A. & Schoeffter, Ph. *Br. J. Pharmac.* **105**, 511–513 (1992).
- Lerman, A. et al. *New Engl. J. Med.* **325**, 997–1001 (1991).
- McMurray, J. J. et al. *Circulation* **85**, 1374–1379 (1992).
- Dohi, Y. & Lüscher, T. F. *Hypertension* **18**, 543–549 (1991).
- Fischli, W., Clozel, M. & Guilly, C. *Life Sci.* **44**, 1429–1436 (1989).
- Arunlakshana, O. & Schild, H. O. *Br. J. Pharmac. Chemother.* **14**, 48–58 (1959).
- Williams, D. L. Jr, Jones, K. L., Pettibone, D. J., Lis, E. V. & Clineschmidt, B. V. *Biochem. biophys. Res. Commun.* **175**, 556–561 (1991).
- Clozel, M., Löffler, B.-M. & Gloor, H. *J. Cardiovasc. Pharmac.* **17**(suppl. 7), S313–S315 (1991).
- Mino, N. et al. *Eur. J. Pharmac.* **221**, 77–83 (1992).
- Webb, M. L. et al. *Biochem. biophys. Res. Commun.* **185**, 887–892 (1992).
- Clozel, M. & Watanabe, H. *Life Sci.* **52**, 825–834 (1993).
- Solomon, R. A. et al. *Stroke* **16**, 58–64 (1985).
- Suzuki, R. et al. *J. Neurosurg.* **77**, 96–100 (1992).
- Clozel, M. *J. Hypertension* **7**, 913–917 (1989).
- Mulvany, M. J. & Halpern, W. *Circ. Res.* **41**, 19–26 (1977).
- Heymann, M. A. & Payne, B. D., Hoffman, J. I. E. & Rudolph, A. M. *Progr. cardiovasc. Dis.* **20**, 55–79 (1977).
- Clozel, J. P., Fischli, W. & Ménard, J. J. *Hypertension* **11**, 75–81 (1993).

Protein composition determines the anti-atherogenic properties of HDL in transgenic mice

Joshua R. Schultz, Judy G. Verstuyft,
Elaine L. Gong, Alex V. Nichols
& Edward M. Rubin*

Life Sciences Division, Lawrence Berkeley Laboratory,
University of California, Berkeley, California 94720, USA

HIGH-DENSITY lipoprotein (HDL) contains two major proteins, apolipoprotein A-I (apoA-I) and apolipoprotein A-II (apoA-II), comprising about 70% and 20% of the total HDL protein mass, respectively. HDL exists in human plasma in two main forms, one containing apoA-I with apoA-II (AI/AII-HDL) and another containing apoA-I without apoA-II (AI-HDL). A strong inverse relationship exists between total plasma HDL concentration and atherosclerosis, but the results of studies examining the relationship between AI-HDL and AI/AII-HDL and atherosclerosis have been conflicting¹⁻⁹. To determine whether these two HDL populations have different effects on atherogenesis, human apoA-I (AI) and human apoA-I and apoA-II (AI/AII) transgenic mice were produced in an atherosclerosis-susceptible strain¹⁰⁻¹². Following an atherogenic diet, despite similar total cholesterol and HDL cholesterol concentrations, the area of atherogenic lesions in the AI/AII mice was 15-fold greater than in the AI animals. These studies show that the protein composition of HDL significantly affects its role in atherogenesis and that AI-HDL is more anti-atherogenic than AI/AII-HDL.

The AI transgenic animals were created using an 11-kilobase human apoA-I genomic fragment¹¹. To obtain transgenic mice producing very high levels of human apoA-II, the human apoA-II structural gene was placed under the control of the human apoA-I promoter. To produce the AI/AII transgenic mice, the AII transgenic mice were then bred into a high-expressing human AI transgenic mouse line¹². The plasma apolipoprotein concentrations in the AI and AI/AII transgenic animals were determined (Table 1). The AI and AI/AII transgenic mice had similarly elevated human apoA-I concentrations and only differed by the presence of high concentrations of human apoA-II in the AI/AII transgenic mice^{11,12}. The concentrations of murine apoA-I and murine apoA-II were similar but low in AI and AI/AII transgenic animals compared to the excess of the human apolipoproteins. Total cholesterol and HDL cholesterol (HDL-C) values did not differ significantly between AI and AI/AII transgenic animals when they were fed the same diet (Table 1). The AI and AI/AII transgenic mice had similar HDL-C concentrations after 10 weeks on the chow diet and nearly identical concentrations after the atherogenic diet. Plasma concentrations of total cholesterol and non-HDL-C were also similar in AI and AI/AII transgenic animals both before and after the atherogenic diet.

Murine apoA-II differs significantly from human apoA-II with regard to both primary structure^{13,14} and its effect on HDL size^{12,14}. Furthermore, although apoA-II concentrations correlate positively with HDL levels in mice^{14,15}, a similar association has not been established for apoA-II in humans^{16,17} or in transgenic mice overexpressing human apoA-II (ref. 12). Structurally, murine apoA-II exists as a monomer and human apoA-II exists as a disulphide-linked dimer of two identical polypeptide chains¹⁸. SDS-polyacrylamide gel electrophoresis followed by immunoblotting of non-reduced and reduced plasma lipoprotein samples was used to determine the molecular form of human apoA-II in the AI/AII transgenic animals (Fig. 1). This analysis

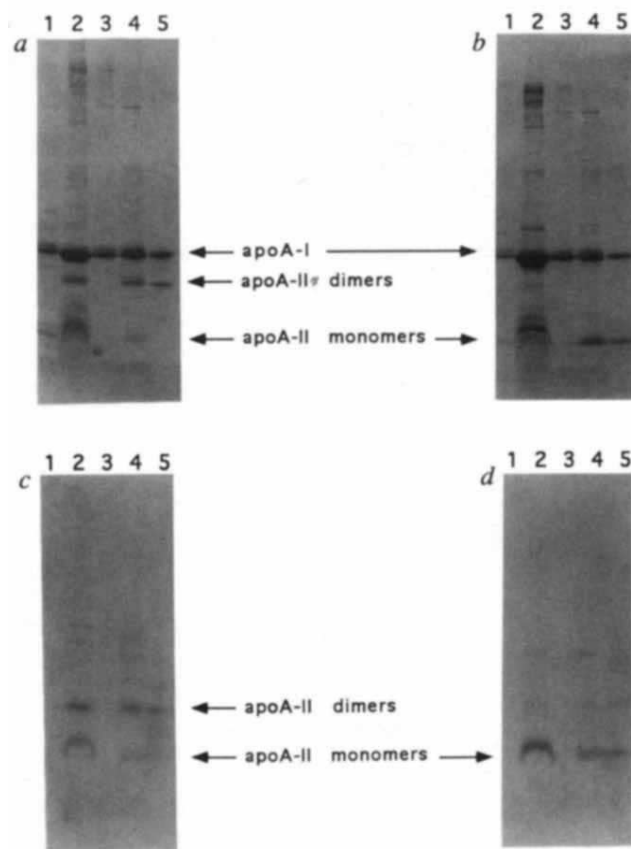


FIG. 1 SDS-PAGE and immunoblot analysis of plasma lipoproteins. Total lipoproteins (of density $\rho \leq 1.21 \text{ g ml}^{-1}$) were isolated from plasma by ultracentrifugation¹². Shown are non-reduced (a, c) and reduced (b, d) samples from (1) C57BL/6 control plasma; (2) human plasma; (3) AI; and (4) AI/AII transgenic mouse plasma, as well as (5) purified human apoA-I and human apoA-II. In c and d, after SDS-PAGE, proteins were electrophoretically transferred to nitrocellulose and allowed to react with goat anti-human apoA-II sera. Binding of antibody to the proteins was visualized by a biotin-conjugated antibody specific for goat IgG and the substrate 4-chloro-1-naphthol.

demonstrated the presence of dimeric human apoA-II in the plasma of the AI/AII transgenic animals (Fig. 1a, c). Reduction by β -mercaptoethanol resulted in the loss of the dimeric apoA-II band and the appearance of a band that co-migrated with monomeric apoA-II (Fig. 1b, d).

The composition of HDL particles from AI/AII, AI and control mice was analysed (Table 2a). Immunoaffinity chromatography was used to separate human apoA-II-containing and human apoA-II-free HDL particles from the plasma of AI/AII transgenic mice¹⁹. Analysis of the separated fractions included determining the concentrations of human and murine apoA-I and apoA-II as well as phospholipid, cholesterol and triglyceride (Table 2a). In the AI/AII transgenic mice about 88% of human apoA-I was associated with HDL particles also containing human apoA-II, indicating a predominance of human AI/AII-HDL. Analysis of the apolipoprotein composition of HDL in AI transgenic mice (Table 2a) demonstrated mainly human AI-HDL particles ($\sim 90\%$). Small amounts of murine apoA-I and apoA-II were also associated with HDL from both groups of transgenic animals (Table 2a).

As apolipoprotein-specific HDL populations from human plasma have distinct size distributions, the size distribution of HDL in the transgenic mice was also determined using non-denaturing gradient gel electrophoresis. HDL from the AI transgenic mice on the chow diet (Fig. 2b) showed a bimodal distribution nearly identical to that characteristic of the AI-HDL

* To whom correspondence should be addressed.

TABLE 1 Apolipoprotein and cholesterol concentrations in AI and AI/AII transgenic and C57BL/6 control mice

Mouse line	Human				Murine				Cholesterol levels					
	apoA-I		apoA-II		apoA-I		apoA-II		Total		HDL		Non-HDL	
	Diet: chow	athero	chow	athero	chow	athero	chow	athero	chow	athero	chow	athero	chow	athero
AI (n=12)	245 (5)	317 (11)	—	—	8 (1)	25 (1)	21 (2)	40 (1)	138 (6)	233 (11)	92 (4)	123 (6)	46 (3)	110 (6)
AI/AII (n=11)	238 (4)	286 (10)	56 (3)	108 (4)	7 (2)	23 (2)	19 (2)	37 (1)	119 (2)	227 (8)	82 (5)	126 (8)	37 (2)	101 (5)
Controls (n=17)	—	—	—	—	110 (2)	135 (2)	24 (1)	46 (2)	96 (2)	168 (10)	59 (2)	54* (3)	37 (2)	114 (10)

Sex- and age-matched transgenic and control mice were housed in the same cages. The values shown represent the mean (and s.e.m.) in mg dl^{-1} for the number of animals (n) shown after a 10-week chow diet and again following 36 weeks of exposure to an atherogenic (athero) diet. The chow diet consists of Purina laboratory mouse chow (5001) containing 4.5% animal fat, <0.03% cholesterol. The athero diet contains 15% fat (primary fat source is dairy butter), 1.25% cholesterol, 0.5% sodium cholate and 20% casein (all percentages are w/w). Values for the plasma concentrations of apoA-I and apoA-II were determined by radial immunodiffusion (RID) assays done with anti-human and anti-mouse apoA-I and apoA-II antibodies, as previously described¹². Total cholesterol was measured in saline with cholesterol reagent (high-performance cholesterol kit; Boehringer-Mannheim). HDL-cholesterol was measured after the removal of non-HDL lipoproteins by selective precipitation with polyethylene glycol. Statistical analysis was done by Student's t-test for paired (chow versus athero diet) and unpaired (AI versus AI/AII transgenic) samples. Differences between AI and AI/AII transgenic animal values were not statistically significant. Differences between chow and athero values were significant at $P < 0.0001$, except that marked with an asterisk, which was not significantly different.

population isolated from human plasma (Fig. 2a). HDL from the AI/AII transgenic mice on the chow diet had a more complex and constrained size distribution (Fig. 2e), similar to the AI/AII-HDL population separated from human plasma (Fig. 2d). After the 36-week atherogenic diet, the size distribution in both AI and AI/AII transgenic mice shifted to one with a higher proportion of particles in the larger HDL size subclasses (Fig. 2c, f).

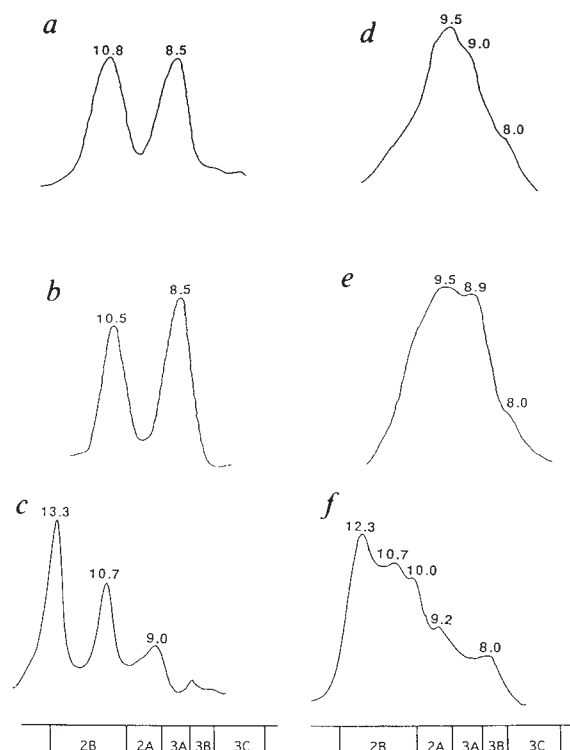


FIG. 2 Non-denaturing gradient gel electrophoresis of human AI-HDL (a) and human AI/AII-HDL (d) populations isolated by immunoaffinity chromatography¹⁹, and of lipoproteins isolated from plasma of AI (b (chow) and c (atherogenic)) and AI/AII (e (chow) and f (atherogenic)) transgenic mice after the diets indicated. Plasma lipoproteins ($p < 1.21 \text{ g ml}^{-1}$) were isolated and samples were run on 4–30% polyacrylamide gels. Computer-assisted scanning densitometry of the Coomassie blue G-250-stained gels was used to determine electrophoretic patterns and particle sizes as described³⁰. The scales beneath the scans indicate the positions of human HDL populations and the numbers above the peaks indicate the particle sizes in nm. HDL population size intervals determined by gradient gel electrophoresis are: HDL_{2b}, 12.9–9.8 nm; HDL_{2a}, 9.8–8.8 nm; HDL_{3a}, 8.8–8.2 nm; HDL_{3b}, 8.2–7.8 nm and HDL_{3c}, 7.8–7.2 nm.

TABLE 2 HDL composition and atherosclerosis of AI and AI/AII transgenic and C57BL/6 control mice

(a) HDL composition

	AI/AII*	AI†	C57BL/6‡
% AI/AII-HDL	88	90	0
% AI-HDL	12§	10	100

	Human	AI/AII¶	AI/AII HDL	AI-HDL¶	C57BL/6
apoA-I _{human}	21 (5)	168 (16)	254 (21)	—	—
apoA-II _{human}	0	50 (7)	0	—	—
apoA-I _{mouse}	8 (3)	9 (2)	11 (2)	121 (9)	—
apoA-II _{mouse}	3 (1)	6 (2)	8 (1)	44 (3)	—

Per cent lipid**

	Human	AI/AII	AI	C57BL/6
Phospholipid	45	51	77	68
Cholesterol	48	43	22	30
Triglyceride	7	6	1	2

(b) Atherosclerotic lesion areas††(μm²)

	AI/AII	AI	C57BL/6
	3,432 (1429)	234 (95)	9,381 (2349)

* Immunoaffinity chromatography was done by applying pooled plasma from 11 AI/AII transgenic mice to an anti-human apoA-II-specific column¹⁹. The % AI- and AI/AII-HDL represent the portion of human apoA-I in the unbound and bound fraction, respectively.

† AI-HDL in the AI transgenic mice was determined by differential electroimmunoassay and AI/AII-HDL was determined by enzyme-linked differential antibody immunosorbent assay⁴.

‡ C57BL/6 mouse HDL is composed of AI/AII-HDL particles²⁷. Protein and lipid composition of the C57BL/6 controls was measured after the removal of non-HDL lipoproteins by selective precipitation with polyethylene glycol²⁸.

§|| Owing to immunological crossreactivity (the anti-human apoA-I immunoaffinity chromatography column retained mouse apoA-I), we could not completely resolve human AI-HDL particles from the AI/AII transgenic mouse plasma. Therefore, the values shown represent the unbound or human AI-free-HDL fraction which is predominantly human AI-HDL and includes (||) small amounts of murine and/or hybrid AI/AII-HDL (that is, with mouse and/or human apoA-I). Protein and lipid concentrations were determined after the removal of non-HDL lipoproteins by selective precipitation with PEG.

¶ The apoA-I and apoA-II concentrations ($\text{mg dl}^{-1} \pm \text{s.e.m.}$) were determined by radial immunodiffusion assays with anti-mouse and anti-human antibodies as previously described¹². No crossreactivity was encountered in radial immunodiffusion assays. Values are the mean (and s.e.m.) of 3 assays on pools of 17 C57BL/6 controls, 12 AI and 11 AI/AII transgenic animals.

** Phospholipid was quantified by the method of Bartlett²⁹. Total cholesterol in the HDL populations was determined as described in Table 1 legend. Triglyceride concentrations were measured using a commercially available enzymatic kit (Sigma Diagnostics) and corrected for free glycerol as previously described¹². Lipid percentages are based on the concentrations (weight) measured.

†† Mean area of atherosclerotic lesions per section per animal²⁰ for 12 AI and 11 AI/AII transgenic and 17 C57BL/6 control animals fed the atherogenic diet for 36 weeks. The difference between AI and AI/AII transgenic mice was significant at $P < 0.007$ by Mann-Whitney U-test analysis.

Control mouse HDL has been previously shown to have a unimodal distribution which shifts to a larger size range when the animals are fed the atherogenic diet²⁰. The size distribution of very-low-density and low-density lipoprotein particles (also measured by non-denaturing gradient gel electrophoresis) did not differ significantly between AI and AI/AII transgenic mice (not shown). The HDL size and apolipoprotein composition data indicate that the HDL particles in the AI and AI/AII transgenic mice no longer resemble those of the mouse but instead share features with the AI- and AI/AII-HDL populations present in human plasma, respectively.

To determine whether the protein composition of HDL affects atherogenesis, the area of lipid-stained lesions was quantified in a defined segment of the proximal aorta (Table 2b; ref. 20). This analysis showed that the mean lipid-staining lesion area per mouse was 234 μm^2 for the AI transgenic and 3,413 μm^2 for AI/AII transgenic mice. Thus, AI/AII transgenic animals had about 15-fold more lesion area per section than AI mice ($P < 0.007$). Parallel analysis of lesion development in 17 non-transgenic C57BL/6 control animals (with HDL-cholesterol levels about half those of the AI and AI/AII transgenic mice) resulted in an average mean lesion area of 9,381 μm^2 . This is 2.7-fold greater than the lesion area in the AI/AII transgenic animals ($P < 0.02$) and more than 40-fold greater than that in the AI transgenic mice ($P < 0.0001$). In a previous study, when C57BL/6J mice were fed an atherogenic diet for 35 weeks, fatty streak aortic lesions progressed to fibro-fatty lesions having many features of advanced human atheromatous plaques, including the presence of lipid-filled macrophages and smooth muscle cells, necrotic debris and cholesterol crystals²¹. In the present study, large lesions from both the AI/AII and non-transgenic C57BL/6 animals fed an atherogenic diet for 36 weeks showed extensive lipid-staining material infiltrating the disrupted media of the vessel wall. In contrast, the limited number of fatty streak lesions from the AI transgenic mice were smaller in area and did not extend into the media.

The composition and size analysis of HDL particles presented in this study indicate that, to a large extent, the two major HDL populations normally found mixed in human plasma have been separated, *in vivo*, in the AI and AI/AII transgenic mice. The difference in susceptibility to atherogenesis between these animals is primarily attributable to the human apoA-II. The presence of apoA-II on HDL containing human apoA-I effectively diminishes the anti-atherogenic effect of AI-HDL. The biochemical mechanism underlying this observation is not known. A hypothetical process, reverse cholesterol transport^{22,23}, remains the focus of many studies aimed at understanding the protective nature of HDL. According to this model, circulating HDL accepts cholesterol from peripheral tissues and promotes its transport to the liver. The individual steps by which HDL may function in this pathway are unclear. A number of *in vitro* studies, but not all, have suggested that AI-HDL is more effective in mediating cholesterol efflux than is AI/AII-HDL²⁴⁻²⁶. The *in vivo* finding of decreased susceptibility to atherogenesis in AI compared to AI/AII transgenic mice with nearly identical lipoprotein profiles is consistent with these findings and provides additional evidence to support the hypothesis that AI-HDL has more anti-atherogenic properties than AI/AII-HDL. The animals described in this report should be useful in future studies aimed at investigating the specific *in vivo* mechanism(s) responsible for the protective role of HDL against atherogenesis and explaining the metabolic role of apoA-II. □

Received 15 June; accepted 16 August 1993.

- Genest, J. J. *et al.* *Circulation* **82**, 621 (1990).
- Puchois, P. *et al.* *Atherosclerosis* **68**, 35-40 (1987).
- Luc, G. *et al.* *Metabolism* **40**, 1238-1243 (1991).
- Fruchart, J. C. & Ailhaud, G. *Clin. Chem.* **38**, 793-797 (1992).
- Lussier-Cacan, S. *et al.* *Atherosclerosis* **78**, 167-182 (1989).
- Valimaki, M. *et al.* *Metabolism* **40**, 1168-1172 (1991).
- Stampfer, M. J., Sacks, F. M., Salvini, S., Willett, W. C. & Hennekens, C. H. *New Engl. J. Med.* **325**, 375-381 (1991).

- Coste-Burel, M., Mainard, F., Chivot, L., Auget, J. L. & Madec, Y. *Clin. Chem.* **36**, 1889-1891 (1990).
- Cheung, M. C., Brown, B. G., Wolf, A. C. & Albers, J. J. *J. Lipid Res.* **32**, 383-394 (1991).
- Paigen, B., Ishida, B. Y., Verstuyft, J. R., Winters, R. & Albee, D. *Arteriosclerosis* **10**, 316-323 (1990).
- Rubin, E. M., Ishida, B. Y., Clift, S. M. & Krauss, R. M. *Proc. natn. Acad. Sci. U.S.A.* **88**, 434-438 (1991).
- Schultz, J. R. *et al.* *J. Biol. Chem.* **267**, 21630-21636 (1992).
- Knott, T. J., Priestley, L. M., Urdea, M. & Scott, J. *Biochem. biophys. Res. Commun.* **120**(3), 734-743 (1984).
- Doolittle, M. H. *et al.* *J. Biol. Chem.* **265**(27), 16380-16388 (1990).
- Mehrabian, M. *et al.* *Arteriosclerosis and Thrombosis* **13**(1), 1-9, (1993).
- Schaefer, E. J., Heaton, W. H., Wetzel, M. G. & Brewer, H. B. *Arteriosclerosis* **2**, 16-26 (1982).
- Civeira, F. *et al.* *Atherosclerosis* **92**, 165-176, (1992).
- Brewer, H. B., Lux, S. E., Ronan, R. & John, K. M. *Proc. natn. Acad. Sci. U.S.A.* **69**, 1304-1308 (1972).
- Cheung, M. C. & Albers, J. J. *J. Biol. Chem.* **259**, 12201-12209 (1984).
- Rubin, E. M., Krauss, R. M., Spangler, E. A., Verstuyft, J. G. & Clift, S. M. *Nature* **353**, 265-267 (1991).
- Stewart-Phillips, J. & Lough, J. *Atherosclerosis* **90**, 211-218 (1991).
- Glomset, J. A. *J. Lipid Res.* **9**, 155-164 (1968).
- Miller, N. E., La Ville, A. & Crook, D. *Nature* **314**, 109-111 (1985).
- Castro, G. R. & Fielding, C. J. *Biochemistry* **27**, 25-29 (1988).
- Barbaras, R., Puchois, P., Fruchart, J. C. & Ailhaud, G. *Biochem. biophys. Res. Commun.* **142**, 63-69 (1987).
- Johnson, W. J., Kilsdonk, E. P. C., van Tol, A., Philips, M. C. & Rothblat, G. H. *J. Lipid Res.* **32**, 1993-2000, (1991).
- LeBoeuf, R. C. *et al.* *J. Lipid Res.* **31**, 91-101 (1990).
- Izzo, C., *et al.* *Clin. Chem.* **27**, 371-374, (1981).
- Bartlett, G. R. *J. Biol. Chem.* **234**, 466-468, (1959).
- Nichols, A. V., Krauss, R. M. & Muslin, T. A. in *Methods in Enzymology: Plasma Lipoproteins* (eds Segrest, J. & Albers, J.) 417-431 (Academic, New York, 1986).

ACKNOWLEDGEMENTS. We thank J.-C. Fruchart for doing the enzyme-linked differential-antibody immunosorbent assays to quantify the two different apoA-I-containing lipoprotein particles, P. Blanche, C. Brown, S. Clift, D. Kumar, E. Spangler, L. Taat, M. Shoukry and C. Chavez for technical assistance, and R. Krauss and L. Stoltzfus for reviewing the manuscript. This research was funded by the National Dairy Promotion and Research Board and administered in cooperation with the National Dairy Council (E.M.R.), by the NIH, National Heart, Lung and Blood Institute (E.M.R.), by the NIH (to E.G. and A.V.N.) and by the American Heart Association (J.R.S.). E.M.R. is an Established Investigator of the American Heart Association. Research was conducted at the Lawrence Berkeley Laboratory through the United States Department of Energy and the University of California, Berkeley.

Tumour-suppressor activity of H19 RNA

Yue Hao*, Taria Crenshaw*, Thomas Moulton†, Elizabeth Newcomb‡ & Benjamin Tycko*§

* Divisions of Oncology and Neuropathology, Department of Pathology, and † Division of Pediatric Hematology-Oncology, Department of Pediatrics, Columbia University College of Physicians and Surgeons, New York, New York 10032, USA

‡ Department of Pathology, New York University School of Medicine, New York, New York 10016, USA

Loss of heterozygosity in certain human embryonal tumours implicates a tumour-suppressor gene at chromosome 11p15.5 and selective loss of maternal alleles suggests that this gene is paternally imprinted¹⁻⁴. The human *H19* gene maps to 11p15.5, is expressed in differentiating fetal cells^{5,11} and is paternally imprinted¹²⁻¹⁶. We report here that two embryonal tumour cell lines, RD and G401, showed growth retardation and morphological changes when transfected with an *H19* expression construct. More importantly, clonogenicity in soft agar and tumorigenicity in nude mice were abrogated in the G401-H19 transfectants. In addition to demonstrating its tumour-suppressor potential, this transfection system should help structural and functional studies of the enigmatic *H19* gene.

To test for effects of *H19* expression, we cloned the gene downstream of the metallothionein promoter in the replicating episomal expression vector pMEP4 (Fig. 1). G401 cells derived from a childhood kidney tumour, originally considered to have been a Wilms' tumour¹⁷ but more recently a malignant rhabdoid tumour¹⁸, were chosen as target cells because they become non-tumorigenic on introduction of chromosome 11p15.5-11p14 by

§ To whom correspondence should be addressed.

FIG. 1 pMEP-H19 DNA and *H19* RNA in transfectants: a, Southern blot of DNA from *H19* and vector control G401 transfectants, hybridized to an *H19* first-exon probe. A diagnostic 0.8-kb *RsaI* fragment derived from pMEP-H19 is present in the G401-H19 cells and absent from the control transfectants. Endogenous *H19* DNA is detected in a band at 2.1 kb and the band at 0.4 kb is generated both by endogenous and pMEP-H19 DNA. Lanes: 1, G401-pMEP control (line 1); 2, G401-H19 (line 3); 3, G401-H19 (line 4); 4, G401-H19 (line 5). Densitometry, with values normalized to the 2.1-kb

bands, indicated episome copy numbers of between four and five per cell in G401-H19 lines 3 and 5 and less than two per cell in G401-H19 line 4. b, Northern analysis of RNA from transfectants and normal human tissues. The *H19* first exon probe detects the predicted full-length transcript at about 2.3 kb (arrowhead) and higher-*M_r* splicing intermediates in the pMEP-H19 lines 3 and 5 and in normal fetal tissues. pMEP-H19 line 4 shows drastically reduced *H19* expression. Hybridization to a β -actin probe is shown as a control for RNA loading and was used to normalize the densitometric data; fetal kidney and placenta lanes on the right were loaded with twice as much RNA to show the rough linearity of the signals. Positions of 28S and 18S ribosomal RNAs are indicated by dashes. KI, Kidney; PI, placenta.

METHODS. pMEP-H19A, used in the initial series of transfections, was made by ligating a 6-kb *EcoRI*-*EcoRI* cloned genomic fragment, containing the entire *H19* gene starting 48 bp upstream of the transcriptional start site, into the *EcoRI* site of pVL1392 (Invitrogen, San Diego), excising the insert with *Bam*HI + *Not*I and ligating it into pMEP4 (Invitrogen). Correct orientation of the insert relative to the metallothioneine promoter was confirmed by restriction analysis. pMEP-H19B, used in later transfections, was made by inserting a 3-kb *Sma*I-*Alu*I partial digestion fragment of the *H19* genomic clone, extending from 15 bp upstream of the transcriptional start site to 0.5 kb downstream of the polyadenylation signal, first into the *Sma*I site of pBluescript KS (Stratagene, La Jolla) and then into the *Kpn*I and *Not*I sites of pMEP4. G401-H19 lines 1 and 2 contained pMEP-H19A; lines 3-9 contained pMEP-H19B. SiHa-H19 cells contained pMEP-H19A. RD-H19 cells contained pMEP-H19B. Cells were transfected using Lipofectin (BRL, Bethesda, MD), selected in $200 \mu\text{g ml}^{-1}$ hygromycin for 3 days and then maintained in $100 \mu\text{g ml}^{-1}$ hygromycin. Individual lines were derived either by pooling 6-10 colonies from a single primary plate (G401 control lines 1 and 2, G401-H19 lines 1-3, SiHa control and SiHa-H19 lines) or by isolating single primary colonies (all other G401 lines and all RD lines). For Southern blotting, $10 \mu\text{g}$ total cellular DNA was digested with *Rsa*I and electrophoresed through 1.2% agarose gels. In the case of the RD transfectants, the presence of the pMEP-H19B episome was confirmed by polymerase chain reaction of DNA from first passage cells using a downstream *H19* first exon primer¹⁵ and an upstream primer matching the polylinker sequence of pMEP-4. For northern blotting, total RNA was prepared by a rapid guanidinium/phenol method²⁴ and 2-10 μg was denatured and electrophoresed through 1.0% agarose gels containing formaldehyde. Densitometry was done on lightly exposed autoradiograms using a Macintosh flatbed scanner (ScannerOne) and Image 1.43 software (NIH). The *H19* first exon probe was generated by genomic PCR as described¹⁵. The actin probe was generated by cDNA PCR using β -actin primers (Clontech, Palo Alto, CA).

microcell fusion¹⁹ and because northern analysis indicated very low expression of the endogenous *H19* gene.

After propagation of transfectants, Southern blotting showed that most lines carried four or five episomes per cell (Fig. 1a). Even without zinc induction, eight of nine lines containing pMEP-H19 expressed at least 10-fold more *H19* RNA than the parental line and vector control transfectants (Fig. 1b). This level of expression was well within the physiological range, lower than in placenta and 10- to 20-fold lower than in 20-week fetal kidney (Fig. 1b). Addition of Zn^{2+} caused an increase in RNA expression in all but one of the G401-H19 lines (Fig. 1b). G401-H19 cells grew readily in culture, but generally at a slower rate

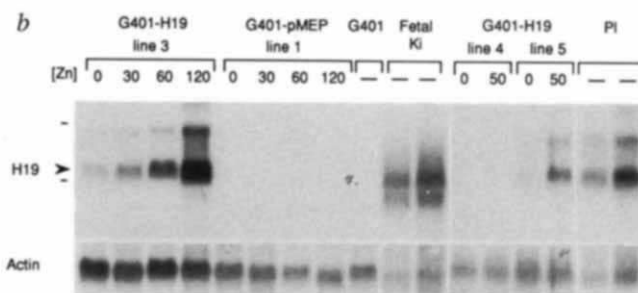
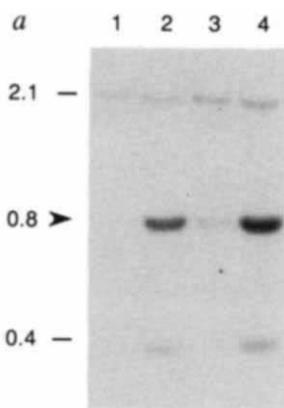
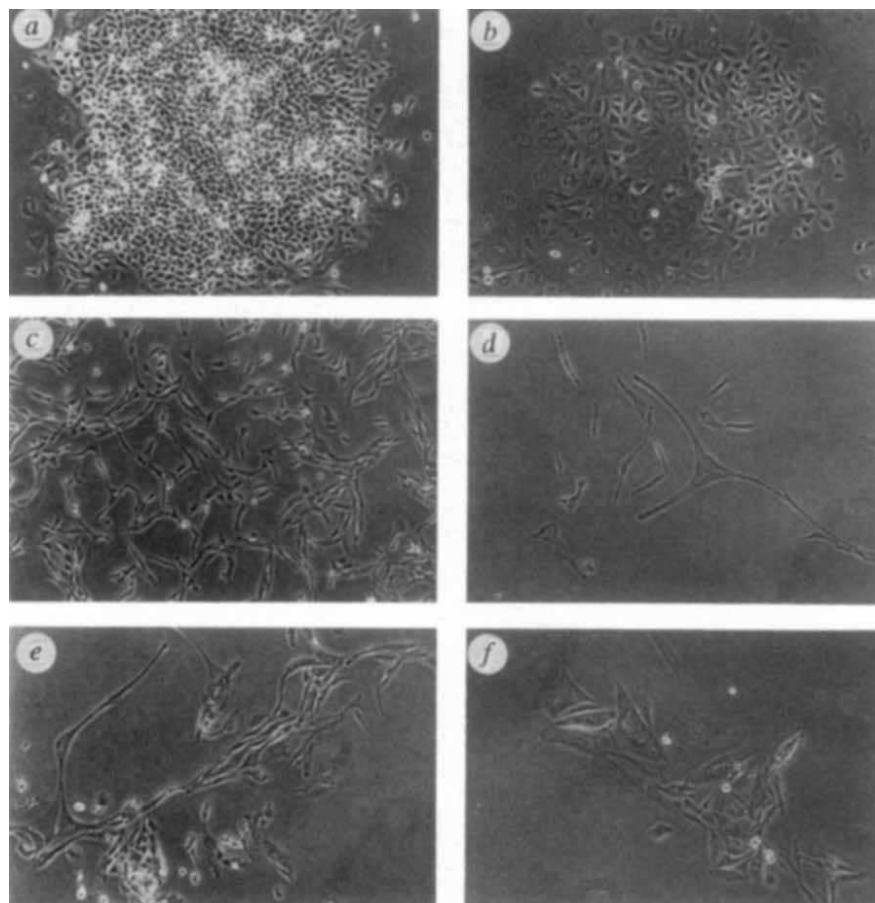


TABLE 1 Soft agar clonogenicity and *in vivo* tumorigenicity of *H19* transfectants

(a)								
Cell line	[Zn ²⁺] (μM)	Soft agar colonies						
SiHa-pMEP	0	2,280						
	50	1,090						
SiHa-pMEP-H19	0	5,480						
	50	1,090						
G401-pMEP		p.3,4	p.5,6	p.7,8	p.10,11	p.12,13	p.15,16	
line 1	0	105 (15)	59 (18)	88 (12)	106 (21)	170 (36)	90 (16)	
line 2	0	303 (22)						
line 3	0	138 (30)						
line 4	0	63 (10)	56 (18)	34 (7)				
line 5	0	18 (4)						
line 1	50	40 (12)	165 (25)	63 (2)	80 (16)	177 (40)	113 (25)	
line 2	50	158 (20)						
line 3	50	170 (20)						
line 4	50	71 (6)	76 (16)	82 (23)				
line 5	50	27 (5)						
G401-pMEP-H19								
line 1	0	0 (0)						
line 2	0	6 (3)	12 (3)					
line 3	0	1 (1)	1 (1)	0 (0)	0 (0)	1 (1)		
line 4	0	14 (2)	17 (4)	6 (1)	11 (2)			
line 5	0	4 (2)	0 (0)	0 (0)	0 (0)			
line 6	0	0 (0)	0 (0)	0 (0)	0 (0)			
line 7	0	1 (1)	0 (0)	0 (0)	0 (0)			
line 8	0	0 (0)	0 (0)	0 (0)	0 (0)			
line 9	0	2 (0)	1 (1)					
line 1	50	1 (1)	0 (0)					
line 2	50	0 (0)	0 (0)					
line 3	50	0 (1)	1 (1)	0 (0)	2 (2)	2 (3)		
line 4	50	24 (4)	20 (4)	18 (6)	13 (1)			
line 5	50	3 (1)	0 (0)	2 (1)	1 (1)			
line 6	50	0 (0)	0 (0)	0 (0)	0 (0)			
line 7	50	1 (1)	0 (0)	0 (0)	0 (0)			
line 8	50	0 (0)	0 (0)	0 (0)	0 (0)			
line 9	50	3 (1)	0 (0)	0 (1)				
(b) Tumorigenicity								
Cell line	Number of tumours/number of injection sites							
G401-pMEP(line 1)					6/6			
G401-pMEP-H19(line1)					0/6			

a, Transfectants were plated in the absence of hygromycin in 0.3% agar at 6×10^3 cells per 35-mm plate and colonies were counted at 2 weeks. Plates were re-examined at 3 weeks, at which time no additional colonies were seen. Values for SiHa cells are the mean for two plates. Values for G401 cells are the mean ($\pm$ s.e.) for three plates. Data from multiple experiments at various passage numbers (p.) are shown: clonogenicity was stable for up to 10 passages. b, G401-H19 (line 1) and G401 control (line 1) transfectants were injected subcutaneously into nude mice. Each mouse received two injections of 3.8×10^6 cells, one on each flank. Tumour formation was assessed grossly and microscopically after 48 days. Tumours formed by the control transfectants arose after a latency of 15-20 days and were from 4 to 10 mm in diameter after 48 days. Microscopically these consisted of undifferentiated cytologically malignant cells with tumour vascularization and focal necrosis. Injection sites which received G401-H19 cells showed only histologically normal skin and subcutaneous tissues, with no residual foreign cells.

FIG. 2 Growth and morphology of transfectants. G401 and RD transfectants were plated at low initial density and individual colonies were photographed after 10 days. *a*, G401 vector control (line 1). *b*, G401-H19 (line 1). *c*, RD vector control (line 2, first passage). The borders of the colony are beyond the field. This morphology and growth rate were observed in the parental line and seven control lines. *d*, *e*, *f*, RD-H19 (three independent lines, first passage). This morphology and growth rate were seen in four of seven RD-H19 lines during the first 5 weeks after transfection. The other three lines grew at the control rate. Scale of enlargement is identical in all fields.



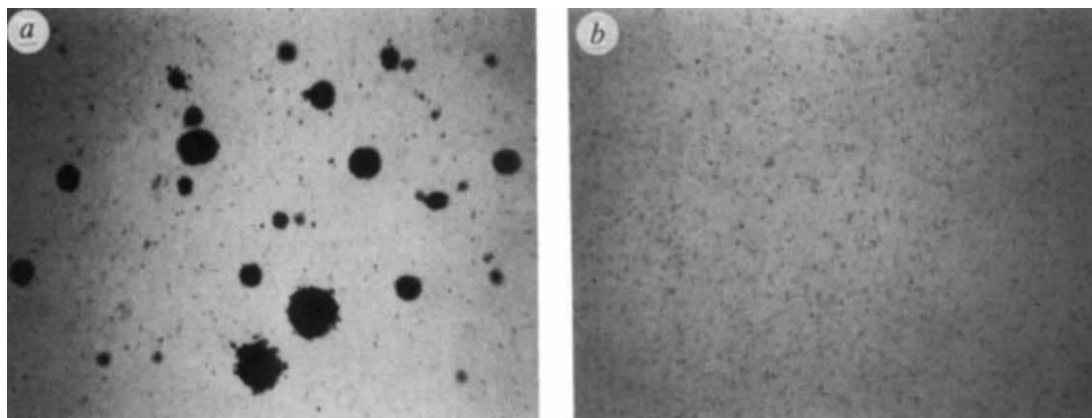
than the vector control transfectants. In the absence of zinc the doubling time of seven of the nine G401-H19 lines was about two days, whereas that of each of the five control lines was less than one day (Fig. 2*a, b*, and data not shown). G401-H19 lines 3 and 4 were exceptions and showed a growth rate equal to that of the control lines. Each of the G401-H19 lines except lines 3 and 4 reached saturation at a cell density of less than half that of the control transfectants and showed a flatter cellular morphology (Fig. 2*a, b*, and data not shown).

To test for anchorage-independent growth, we plated the G401 transfectants in soft agar. Although cells from each of the five vector control lines formed numerous colonies both in the absence and presence of zinc, eight of the nine G401-H19 lines

showed a marked reduction in colony formation even in the absence of zinc (Table 1*a* and Fig. 3). G401-H19 line-2 cells produced a few colonies in the absence of zinc but showed abrogation of clonogenicity in 50 μM Zn^{2+} (Table 1*a*). Clonogenicity was suppressed even in the rapidly growing G401-H19 line 3, indicating a dissociation from growth rate like that seen previously in microcell transfers²⁰. We next assessed the ability of the transfectants to form tumours *in vivo*: when G401 vector control cells and G401-H19 cells were injected subcutaneously into nude mice, the control cells formed tumours whereas the H19 transfectants were non-tumorigenic (Table 1*b*).

Among the G401-H19 lines, the single exception to suppression of clonogenicity was seen in line 4. These cells formed

FIG. 3 Soft agar colony formation by G401 transfectants. Clonogenicity was assayed as described in Table 1. *a*, Low-power ($\times 2.8$) photomicrograph of colonies formed by G401 vector control (line 2) cells. *b*, Absence of colonies in a comparable field from a plate containing G401-H19 (line 2) transfectants. No Zn^{2+} was added in this experiment.



colonies in soft agar (though at a lower efficiency than the vector control cells), even in the presence of zinc. The findings with this line supported a direct relationship between suppression of clonogenicity and *H19* expression. First, Southern blotting showed an episome copy number of less than two per cell, from two- to threefold less than that seen in the other lines (Fig. 1a). Second, the pMEP-H19 episome in this line was transcriptionally inactive: basal *H19* RNA expression was at least 10-fold lower than that seen in the other lines, with no induction in response to Zn^{2+} (Fig. 1b).

SiHa cells from a cervical carcinoma become non-tumorigenic on introduction of chromosome 11 (ref. 21), and RD cells from an embryonal rhabdomyosarcoma show a variable percentage of growth-retarded colonies on introduction of 11p15 subchromosomal fragments²². Neither line expresses *H19* (data not shown). Transfection of SiHa cells with pMEP-H19 had no apparent phenotypic consequences. In spite of a strong induction of *H19* RNA, SiHa-H19 cells were morphologically unaltered, grew at the same rate as control transfectants and remained highly clonogenic in soft agar (Table 1, and data not shown). In contrast, observations in the early post-transfection period indicated that H19 can cause growth suppression in RD cells. In the first five weeks after transfection, four of seven RD-H19 lines grew slowly, with a high frequency of bizarre multinucleate giant cells, features not seen in the parental line or in seven vector control lines (Fig. 2c-f).

H19 was originally cloned in screenings aimed at isolating genes whose transcription increases with cellular differentiation⁵⁻⁷ and it is expressed as an abundant spliced and polyadenylated RNA in a wide array of fetal tissues, including the precursor tissues of Wilms' tumour, malignant rhabdoid tumour and embryonal rhabdomyosarcoma⁵⁻¹¹. Consistent with a potential tumour-suppressor role, *H19* RNA was undetectable by northern blotting in two primary Wilms' tumours that had lost the maternal allele and was variably reduced from fetal kidney levels in three cases which retained both alleles (data not shown). A mechanistic explanation for the antitumorigenic effect of *H19* awaits an understanding of its normal function. The *H19* transcript appears to contain no long translational reading frames and the short open reading frames are not evolutionarily conserved²³. Because *H19* RNA is associated with non-polyosomal proteins in the cytoplasm, it has been suggested that the gene may function directly as the RNA component of a ribonucleoprotein²³. If this is so, then *H19* will define a new class of tumour-suppressor gene. Transfection of modified *H19* constructs into G401 cells may be a convenient way to examine the structural requirements for *H19* function. □

Received 16 June; accepted 17 August 1993.

- Schroeder, W. T. et al. *Am. J. hum. Genet.* **40**, 413-420 (1987).
- Williams, J. C., Brown, K. W., Mott, M. G. & Maitland, N. J. *Lancet* **i**, 283-284 (1989).
- Pal, N. et al. *Oncogene* **5**, 1665-1668 (1990).
- Scrabble, H. et al. *Proc. natn. Acad. Sci. U.S.A.* **86**, 7480-7484 (1989).
- Davis, R. L., Weintraub, H. & Lassar, A. B. *Cell* **51**, 987-1000 (1987).
- Pachnis, V., Brannan, C. I. & Tilghman, S. M. *EMBO J.* **7**, 673-681 (1988).
- Wiles, M. V. *Development* **104**, 403-413 (1988).
- Poirier, F. et al. *Development* **113**, 1105-1114 (1991).
- Brunkow, M. E. & Tilghman, S. M. *Genes Dev.* **5**, 1092-1101 (1991).
- Rachmilewitz, J. et al. *Molec. Reprod. Dev.* **32**, 196-202 (1992).
- Han, D. K. & Liao, G. *Circulation Res.* **71**, 711-719 (1992).
- Bartolomei, M. S., Zemel, S. & Tilghman, S. *Nature* **351**, 153-155 (1991).
- Zhang, Y. & Tycko, B. *Nature Genet.* **1**, 40-44 (1992).
- Rachmilewitz, J. et al. *FEBS Lett.* **309**, 25-28 (1992).
- Zhang, Y. et al. *Am. J. hum. Genet.* **53**, 113-124 (1993).
- Rainier, S. et al. *Nature* **362**, 747-749 (1993).
- Peebles, T. T., Trisch, T. & Papageorge, A. G. *Pediatr. Res.* **12**, 485 (1978).
- Garvin, A. J., Re, G. G., Tarnowski, B. I., Hazen-Martin, D. J. & Sens, D. A. *Am. J. Path.* **142**, 375-380 (1992).
- Dowdy, S. F. et al. *Science* **254**, 293-295 (1991).
- Weissman, B. E. et al. *Science* **236**, 175-180 (1987).
- Koi, M. et al. *Molec. Carcinogen.* **2**, 12-21 (1989).
- Koi, M. et al. *Science* **260**, 361-364 (1993).
- Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. *Molec. cell. Biol.* **10**, 28-36 (1990).
- Chomczynski, P. & Sacchi, N. *Analyt. Biochem.* **162**, 156-159 (1987).

ACKNOWLEDGEMENTS. We thank K. Cechova for technical advice and R. Dalla-Favera for discussions. This research was supported by a grant from the NIH.

Distinct NF- κ B/Rel transcription factors are responsible for tissue-specific and inducible gene activation

Thomas Lernbecher, Urs Müller* & Thomas Wirth†

Zentrum für Molekulare Biologie Heidelberg, Im Neuenheimer Feld 282, D-6900 Heidelberg, Germany

* Basel Institute for Immunology, Grenzacherstrasse 487, CH-4005 Basel, Switzerland

THE NF- κ B/Rel family is a growing class of transcriptional regulators¹⁻¹¹ whose members share the conserved Rel-homology domain, involved in specific DNA binding and dimerization. They interact with the regulatory elements of many different genes and are involved in the regulation of lymphoid-specific and inducible transcription¹². We tested whether these factors could alone activate a gene in transgenic mice. We report here that a minimal promoter containing three copies of a binding site for these proteins allows tissue-specific and inducible transgene activation. In lymphoid tissues constitutive transgene expression correlates with the presence of a constitutively active p50/RelB heterodimer. Other organs that only contain the p50 homodimer do not express the transgene. In contrast to this constitutive activity mediated by p50/RelB, the p50/p65 heterodimer (which is NF- κ B) could confer inducible transgene activation in embryo fibroblasts. Thus two different members of the NF- κ B/Rel family of transcriptional activators are involved in tissue-specific and inducible gene activation in transgenic mice.

Three copies of the κ B motif from the immunoglobulin kappa (Ig κ)-enhancer were inserted immediately upstream of the β -globin TATA box and the β -globin reporter gene (Fig. 1a)^{13,14}. Transient transfection experiments confirmed that this synthetic promoter-enhancer directs high levels of constitutive expression in murine S194 B cells and 12-O-tetradecanoylphorbol 13-acetate (TPA)-inducible expression in murine EL4 T cells (data not shown). Four independent transgenic mouse lines were generated that contained between 4 and 40 copies of the transgene. As in previous studies with transformed cell lines we expected mature B cells in the spleen to contain active nuclear NF- κ B¹⁵. We therefore analysed transgene expression using reverse transcriptase polymerase chain reaction (RT-PCR) with total spleen RNA. Expression was detectable in all four mouse lines (Fig. 1b), whereas no signal could be seen with RNA from a non-transgenic litter (Fig. 1b, lane 7). The absolute levels of expression varied between individual lines, the high expressing lines tg4 and tg5 containing about 30- to 50-fold more transcripts than lines tg9 and tg14. There was no correlation between transgene copy numbers and expression levels (see legend to Fig. 1). The specific transgene integration site is probably responsible for this variation in activity. Nevertheless all mouse lines did show transgene expression indicating that in all cases the κ B motif allowed transgene activation. To confirm that the observed signals originated from the transgene promoter element, total spleen RNA from the tg4 and tg5 lines was analysed by primer extension. The signal for correctly initiated RNA was readily detectable (Fig. 1c).

Transgene expression is only detectable in lymphoid tissues but not in any of the non-lymphoid tissues analysed. The results for line tg5 are shown in Fig. 2a, identical results were obtained for the other mouse lines. In addition, transgene expression is inducible in transgenic embryo fibroblasts. Treatment of these cells with tumour necrosis factor- α (TNF α) resulted in a rapid induction of transgene expression. The signal was detectable

† To whom correspondence should be addressed.

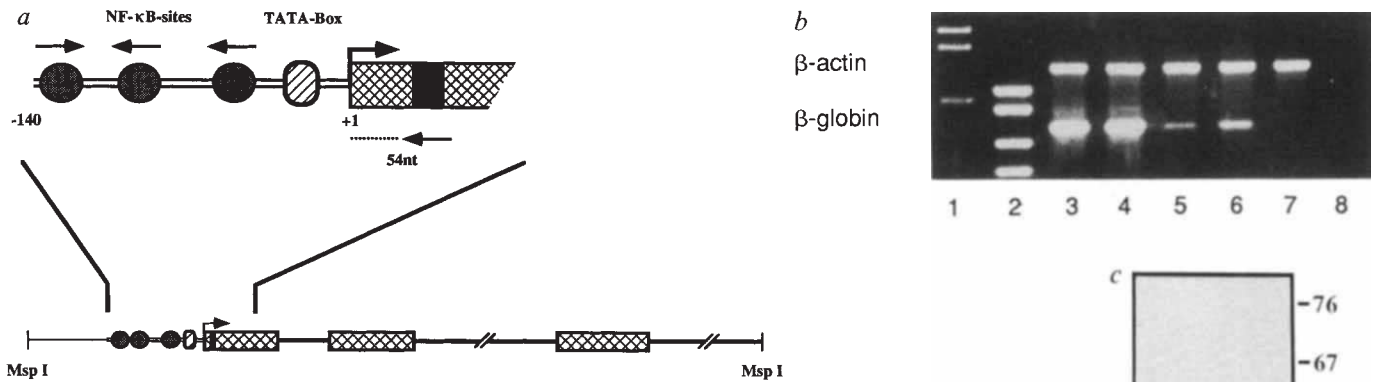


FIG. 1 The transgene is expressed in the spleen of transgenic mice. *a*, Schematic representation of the transgene. The β -globin reporter cassette with a single κ B motif upstream of the globin TATA-box has been described¹⁴. In addition a κ B dimer was inserted 35 base pairs (bp) upstream of the first κ B motif¹³. The orientation of the κ B motifs (stippled circles) with respect to the motif present in the Ig κ enhancer is indicated by arrows. The black box shows the position of a *Clal*-linker insert used to generate transgene-specific primers. The position of the primer used for primer extension and the size of the specific extension product is indicated. The globin exons are shown as hatched boxes, the introns and 3' sequences are drawn as thick lines. The second intron and the 3' flanking region are not drawn to scale. The thin line represents a short piece of pBR322 sequences upstream of the synthetic enhancer/promoter. The total size of the fragment is 2.4 kilobases (kb). *b*, RT-PCR of total spleen RNA with primers specific for the transgene and the endogenous β -actin gene, respectively. Lane 1 shows the 564–947 bp fragments of a Lambda DNA *EcoRI*–*HindIII* digest, lane 2 shows the 309–622 bp fragments of a pBR322 DNA–*MspI* digest, lanes 3 to 6 show the amplified products from the transgenic lines tg4 (lane 3), tg5 (lane 4), tg9 (lane 5) and tg14 (lane 6), respectively. Lane 7 shows the analysis for a non-transgenic litter, no cDNA was added to the amplification reaction in lane 8. The identity of the visualized bands was confirmed by Southern hybridizations. *c*, Primer extension analysis with total spleen RNA from tg4 (lane 1), tg5 (lane 2) and control mice (lane 3). Positions of the 26–76 bp fragments of a pBR322 DNA–*MspI* digest are indicated on the right side of the figure. The upper panel shows the result for the transgene-specific primer (β -globin), the lower panel shows a control with the GAPDH-specific primer.

METHODS *a*, Generation and initial characterization of the transgenic mice was as described²⁶. For estimations of transgene copy numbers the signals from appropriate dilutions of plasmid DNA carrying the transgene were used as reference in Southern blot analyses. The copy numbers are as follows: tg4, 3–5 copies; tg5, 15–20 copies; tg9, 8–10 copies; tg14, 25–30 copies. *b*, cDNA was reverse transcribed from 2 μ g total spleen RNA with oligo dT primers and MuLV reverse transcriptase as described²⁷. 5% of the reaction was amplified in 30 ml for 30 cycles (40 at 94 °C, 1 min 65 °C and 2 min at 72 °C). The following primers were used: 5'-TTTCTGATAGGCAG-CCTGCACTGGT-3' (transgene 5'); 5'-CATAGTTGTGTCAGATCGATCTGG-3' (transgene 3'); 5'-AGAG-GTATCCTGACCCTGAAGTACC-3' (β -actin 5') and 5'-CCACCAGACAACACTGTGTTGGCAT-3' (β -actin 3'). Amplified products were separated on a 1.3% agarose gel and visualized by ethidium bromide staining. The positions of the signals for the β -actin (720 bp) and the β -globin/transgene (440 bp) amplification products are indicated. *c*, Preparation of total RNA and conditions for the primer extension reactions were as described²⁶, except that the globin and GAPDH reactions were done separately.

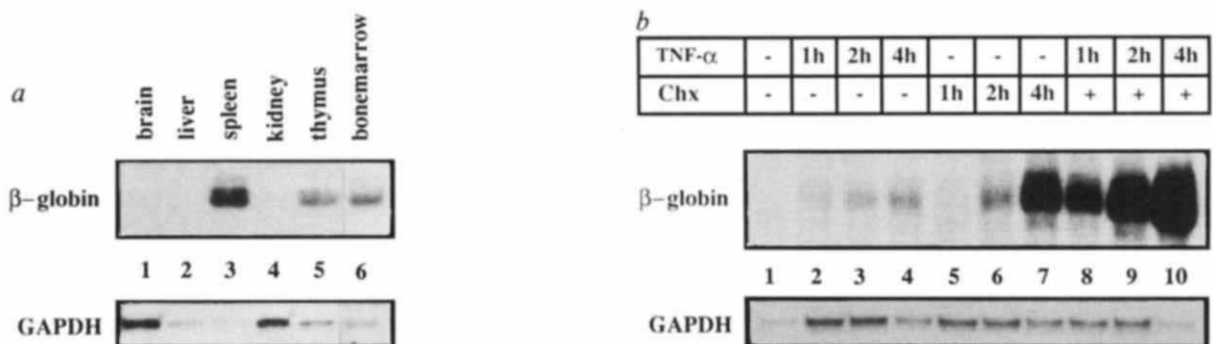


FIG. 2 Transgene expression is restricted to lymphoid tissues and inducible in embryo fibroblasts. *a*, Primer extension analysis of total RNA from the indicated organs of a tg5 mouse. Reaction conditions were as in Fig. 1c. *b*, Primer extension experiment with cytoplasmic RNA from embryo fibroblasts. Treatment of the cells with TNF α (1.8×10^3 U ml⁻¹) and Chx (10μ g ml⁻¹) is indicated above each lane. In the last 3 lanes, cycloheximide was added 20 min before the TNF α induction. Conditions

for primer extension were as above.

METHODS Embryo fibroblasts cells were established from day 14 heterozygous tg5 fetuses essentially as described²⁸ except that they were grown in Iscove's modified DMEM supplemented with 10% FCS. Homogeneous fibroblast cultures had been kept for several months before they were used for the experiments described.

within 1 hour and a plateau was reached after 2 to 4 hours (Fig. 2b, lanes 1 to 4). On long exposures trace amounts of specifically initiated transgene RNA were visible in the untreated cells, probably due to induction by growth factors present in the culture medium¹⁶. Transgene induction was independent of *de novo* protein synthesis, because it could be observed in the presence of cycloheximide (Fig. 2b). In fact, consistent with previous reports¹⁷, inclusion of cycloheximide resulted in a superinduction of transgene expression with TNF α (Fig. 2b, lanes 8 to 10). These results confirm that inducible expression does not require a lymphoid-specific component or newly synthesized proteins.

To correlate the observed transgene expression pattern with κ B-specific transcription factors, we analysed extracts from the various tissues for the presence of κ B binding activity. Nuclear NF- κ B has been described in mature B cell lines, activated T cells and activated macrophages^{15,18,19}. In most other cells NF- κ B is sequestered in the cytoplasm because of the interaction with the I κ B inhibitory proteins but can be induced by various stimuli^{19,20}. Consistent with the tissue-restricted expression (Fig. 2a), strong specific complexes comigrating with the κ B-specific complexes from S194 mouse plasmacytoma cells were detected in spleen and thymus extracts (Fig. 3a, lanes 4 and 6) and in bone

marrow extracts (data not shown). The lower, faster migrating complex could also be detected in the non-lymphoid extracts on longer exposures. The integrity of the different extracts was verified using a probe for octamer transcription factors. Only the complex representing the ubiquitous Oct1 protein is shown.

As different members of the NF- κ B/Rel family can interact with the κ B motif, we used specific antisera to determine the polypeptide composition of the κ B-specific complexes from spleen and thymus. The lower complex is completely abolished when a p50-specific antiserum is included, whereas none of the other antisera affected this complex (Fig. 3b, c). This complex is probably the p50 homodimer which had been identified as KBF1/H2TF1 or NF- κ C in various cell types²¹⁻²³. The p50-specific antiserum also reacted with the upper complex albeit somewhat more weakly, consistent with previous reports showing that p50-specific antisera react better with the homodimer than with heterodimers containing p50^{3,24}. The upper complex did not react with a p65-specific antiserum, indicating that this complex is not genuine p50/p65 NF- κ B. But this complex reacted with the RelB-specific antiserum (Fig. 3b, c). The conclusion that this complex represents a p50/RelB heterodimer was further corroborated by the analyses of complexes generated

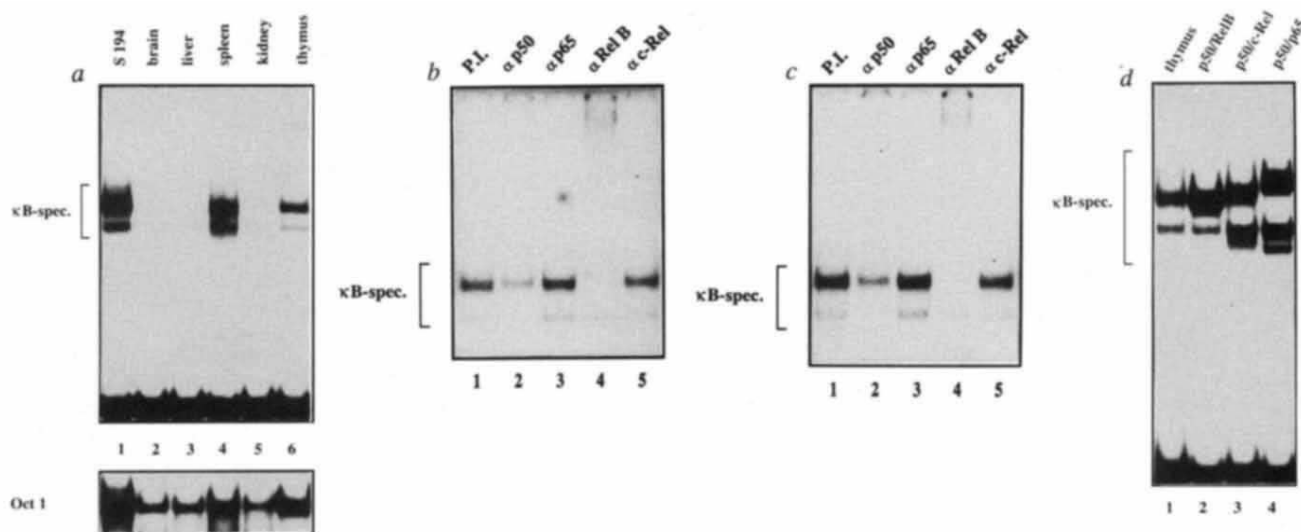


FIG. 3 Extracts from primary lymphoid cells contain a constitutive p50/RelB heterodimer. a, Electrophoretic mobility shift assay (EMSA) with a κ B-specific probe and whole-cell extracts from the organs indicated. As a positive control a nuclear extract from the murine S194 plasmacytoma cell line²⁹ is shown for comparison (lane 1). The position of κ B-specific complex is indicated at the left. The lower part shows an EMSA with an octamer-specific probe. The position of the ubiquitous Oct1 complex is indicated at the left side of the panel. b, c, Characterization of the polypeptide composition of the κ B-specific complexes. The extracts from spleen (b) and thymus (c) were preincubated with either preimmune serum (P.I., lane 1) or rabbit antisera raised against the proteins indicated on top of each lane before addition of the κ B-specific probe. The κ B-specific complexes are indicated. d, The upper κ B-specific complex from thymus comigrates with the recombinant p50/RelB heterodimer. Extracts from thymus or COS1 cells transfected with the combinations of expression vectors indicated on top of each lane were analysed with the κ B-specific probe. e, Splenic B220 positive cells contain the p50/RelB heterodimer. EMSA with the κ B-specific probe and extracts from total spleen (lane 1), the B220 positive fraction (lane 2) and the B220 depleted fraction (lane 3). The position of κ B-specific complexes is indicated.

METHODS. a, Whole-cell extracts were prepared by freezing the organs in liquid nitrogen, homogenization of the frozen tissues on dry ice and then extraction of the homogenate with 2 vols cold buffer C (ref. 30) supplemented with 0.2 mM EGTA by three freeze-thaw cycles before removing insoluble material by a 10-min spin at 4 °C. For EMSA, 10 μ g extract were incubated with 10,000 c.p.m. labelled probe, 3 μ g poly (dI:dC), 10 μ g BSA, in buffer containing 10 mM Tris-HCl pH 7.5, 50 mM NaCl, 1 mM DTT, 1 mM EDTA and 5% glycerol. Complexes were separated on 4% native acrylamide gels run in 0.5 $\times$ TBE. b, c, In the reactions where specific antisera were included, extracts were preincubated with 1–2 μ l rabbit serum for 10 min at 20 °C before the other components were added. d, COS1 cells were transfected with 2.5 μ g of each of the indicated pRc/CMV (Invitrogen) expression vectors by electroporation. The RelB expression vector was generated by PCR amplification of the relB cDNA from S194 cells and cloning into pRc/CMV. e, B220 positive cells were enriched by selection with magnetic beads coated with an anti mouse B220 antibody (Paesel). The first round of selection was used as B220 positive cells, two more depletions were done to generate the B cell-depleted fraction. FACS analysis with antibodies for B-cell-specific (B220, Ig) and T-cell-specific (Thy1) markers confirmed that greater than 90% of the B cells were depleted from the B220 negative fraction.

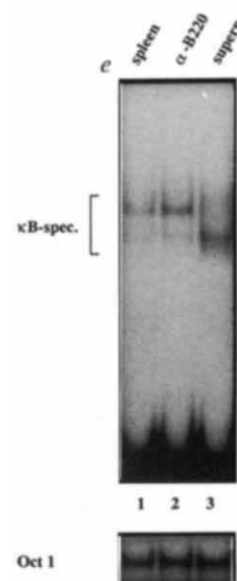
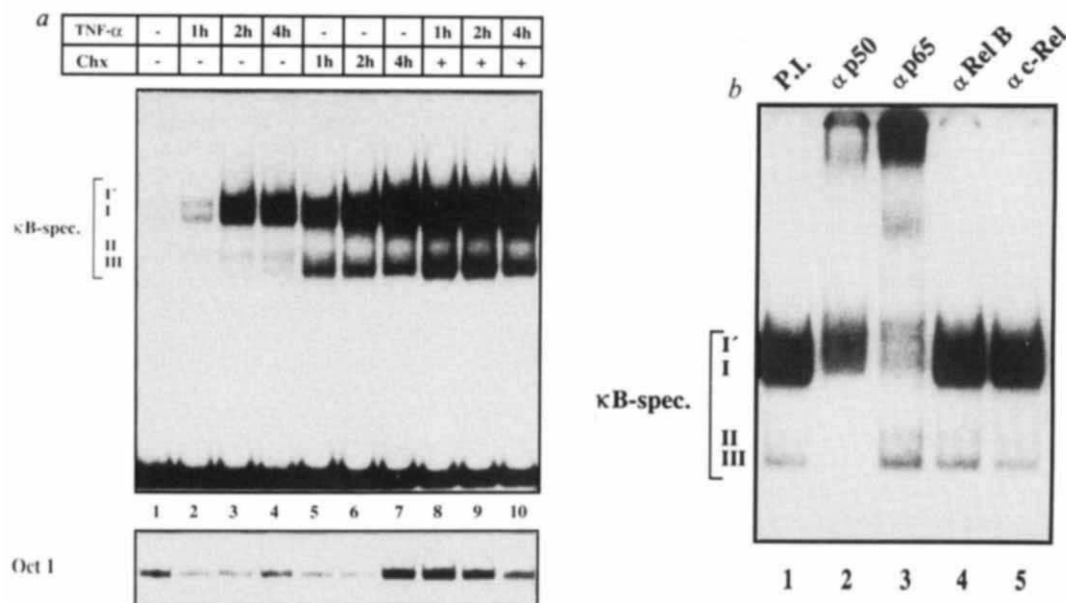


FIG. 4 The inducible p50/p65-heterodimer activates the transgene in TNF α -stimulated fibroblasts. **a**, EMSA with an NF- κ B-specific DNA probe (upper panel) and nuclear extracts from cells that were treated as indicated. The position of the κ B-specific complexes is shown on the left side. The lower panel shows the control shift with the octamer-specific probe. **b**, Characterization of the polypeptide composition of the κ B-specific complexes. The extract corresponding to the 2-h TNF α treatment was analysed with the different antisera indicated on top. Conditions were as described in Fig. 3.



with recombinant Rel-proteins. This complex comigrated with a complex generated by cotransfecting Cos1 cells with expression vectors for p50 and RelB, whereas the p50/p65 and the p50/c-Rel complexes showed a slightly different migration (Fig. 3d). In mock transfected Cos1 cells, a weak complex comigrating with the p50 homodimer is only visible on long exposures (data not shown and ref. 25). The thymus is a primary lymphoid organ that contains predominantly immature T cells. In contrast, the spleen contains mature B and T cells and several other cell types. To identify positively the cell type that contains the p50/RelB complex, we enriched splenic B cells. Figure 3e shows that the p50/RelB complex is present in the B-cell fraction and only barely detectable in the B-cell-depleted fraction. This result suggests that the B cells do contain most of the p50/RelB complex in the spleen. But we cannot completely rule out that a minor fraction of copurifying cells might be the source for the observed complex. Ryseck and colleagues had previously shown that RelB can heterodimerize with p50 and that this heterodimer bound to and transcriptionally activated the κ B-motif¹¹. Our results show that the p50/RelB heterodimer is constitutively active in primary lymphoid cells and its presence correlates with constitutive lymphoid-specific transgene expression. This suggests that it is this complex that is responsible for activation and expression of genes like the Ig κ light chain and others in normal B- and T-cell development. The p50/RelB complex is also constitutively present in several transformed murine B-cell lines: PD31 pre B cells and S194 plasmacytoma cells (data not shown).

Finally, the κ B-specific complexes mediating inducible trans-

gene expression were analysed. Only weak signals are detectable in uninduced cells (Fig. 4a, lane 1). TNF α rapidly induced complexes I' and I, whereas complex II and complex III remained essentially unaffected. The difference between complexes induced in the presence of cycloheximide was that complex II disappeared and complex III was also increased (Fig. 4a, compare lanes 1 to 4 with lanes 5 to 10). Complex II was abolished by the p50-specific antiserum indicating that it represents the p50 homodimer complex (Fig. 4b, lane 2). Complexes I' and I reacted with the p50 and p65-specific antisera indicating that they contain the p50/p65 heterodimer of NF- κ B (Fig. 4b). The reason for the slower migration of complex I' as compared to complex I is not clear. One possible explanation would be that it contains one or more additional protein(s). Complex III is also affected by the p50-specific antiserum and not by any of the other antisera. A complex with similar induction kinetics and mobility has recently been observed in serum-stimulated 3T3 fibroblasts¹⁶. All complexes were specific for the NF- κ B motif and were not observed with a mutant NF- κ B probe (data not shown).

Thus, different members of the family of NF- κ B/Rel family of transcription factors are involved in different cellular transcriptional responses. Apparently, in primary lymphoid cells such as thymocytes and splenic B cells, the p50/RelB heterodimer is constitutively active and therefore most probably involved in transcriptional regulation of κ B-containing promoters and enhancers. In contrast, the p50/p65 heterodimer is inducible in many cell types and confers inducible transcriptional activity on the κ B-motifs in these cells. □

Received 8 June; accepted 2 August 1993.

- Wilhelmsen, K. C., Eggleston, K. & Temin, H. M. *J. Virol.* **52**, 172–182 (1984).
- Steward, R. *Science* **238**, 692–694 (1987).
- Kieran, M. et al. *Cell* **62**, 1007–1018 (1990).
- Gosh, S. et al. *Cell* **62**, 1019–1029 (1990).
- Bours, V., Villalobos, J., Burd, P. R., Kelly, K. & Siebenlist, U. *Nature* **348**, 76–80 (1990).
- Meyer, R. et al. *Proc. natn. Acad. Sci. U.S.A.* **88**, 966–970 (1991).
- Nolan, G. P., Gosh, S., Liou, H.-C., Tempst, P. & Baltimore, D. *Cell* **64**, 961–969 (1991).
- Ruben, S. M. et al. *Science* **251**, 1490–1493 (1991).
- Schmid, R. M., Perkins, N. D., Duckett, C. S., Andrews, P. C. & Nabel, G. J. *Nature* **352**, 733–736 (1991).
- Bours, V. et al. *Molec. cell. Biol.* **12**, 685–695 (1992).
- Ryseck, R.-P. et al. *Molec. cell. Biol.* **12**, 674–684 (1992).
- Lenardo, M. J. & Baltimore, D. *Cell* **58**, 227–229 (1989).
- Pierce, J. W., Lenardo, M. & Baltimore, D. *Proc. natn. Acad. Sci. U.S.A.* **85**, 1482–1486 (1988).
- Wirth, T. & Baltimore, D. *EMBO J.* **7**, 3109–3113 (1988).
- Sen, R. & Baltimore, D. *Cell* **46**, 705–716 (1986).
- Baldwin, A. S. Jr, Azizkhan, J. C., Jensen, D. E., Beg, A. A. & Coodly, L. R. *Molec. cell. Biol.* **11**, 4943–4951 (1991).
- Sen, R. & Baltimore, D. *Cell* **47**, 921–928 (1986).
- Nabel, G. & Baltimore, D. *Nature* **326**, 711–713 (1987).

- Baeuerle, P. A. & Baltimore, D. in *The Hormonal Control Regulation of Gene Transcription* (eds Cohen, P. & Foulkes, J. G.) 423–446 (Elsevier-Biomedical, Amsterdam, 1991).
- Baeuerle, P. A. & Baltimore, D. *Cell* **53**, 211–217 (1988).
- Yano, O. et al. *EMBO J.* **6**, 3317–3324 (1987).
- Baldwin, A. S. Jr & Sharp, P. A. *Proc. natn. Acad. Sci. U.S.A.* **85**, 723–727 (1988).
- Kang, S.-M., Tran, A.-C., Grillo, M. & Lenardo, M. J. *Science* **256**, 1452–1456 (1992).
- Beg, A. A., Finco, T. S., Nantermet, P. V. & Baldwin, A. S. Jr *Molec. cell. Biol.* **13**, 3301–3310 (1993).
- Schmitz, M. L. & Baeuerle, P. A. *EMBO J.* **10**, 3805–3817 (1991).
- Anweiler, A., Müller, U. & Wirth, T. *Nucleic Acids Res.* **20**, 1503–1509 (1992).
- Ausubel, F. M. et al. *Current Protocols in Molecular Biology* (Wiley, New York, 1989).
- Robertson, E. J. in *Teratocarcinomas and Embryonic Stem Cells: a Practical Approach* (ed. Robertson, E. J.) 71–112 (IRL, Oxford, 1987).
- Atchison, M. L. & Perry, R. P. *Cell* **48**, 121–128 (1987).
- Dignam, J. D., Lebovitz, R. M. & Roeder, R. G. *Nucleic Acids Res.* **11**, 1475–1489 (1983).

ACKNOWLEDGEMENTS. We thank A. Buder for help in the initial genotyping of the transgenic animals, A. Israel, M. Lenardo and R. Bravo for the gift of specific rabbit antisera, K. Schulze-Osthoff for TVF α and M. Schmitz and P. Baeuerle for p50, p65 and c-rel expression vectors, J. Afferink, A. Klevenz and B. Arnold for help with the cell separation and FACS analyses and C. Clayton, M. Gossen, H. König and A. Anweiler for reading the manuscript.